AFRL-IF-RS-TR-1999-162
In-House Report
November 1999

SIMULTANEOUS ADAPTIVE CO-CHANNEL SPEAKER SEPARATION

Daniel S. Benincasa

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

Although this report references limited documents (*), listed on pages 184 and 185, no limited information has been extracted.

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-1999-162 has been reviewed and is approved for publication.

APPROVED:

GERALD C. NETHERCOTT
Chief, Multi-Sensor Exploitation Branch
Info & Intel Exploitation Division
Information Directorate

FOR THE DIRECTOR: John V McNamara

JOHN V. MCNAMARA, Tech Advisor
Info & Intel Exploitation Division
Information Directorate

If your address has changed or if you wish to be removed from the Air Force Research Laboratory Rome Research Site mailing list, or if the addressee is no longer employed by your organization, please notify AFRL/IFEC, 32 Brooks Road, Rome, NY 13441-4114. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
November 1999

3. REPORT TYPE AND DATES COVERED
In-House, 1/94 - 1/98

4. TITLE AND SUBTITLE
SIMULTANEOUS ADAPTIVE CO-CHANNEL SPEAKER SEPARATION

5. FUNDING NUMBERS
PE: 62702F
PR: 4594
TA: 15
WU: F0

6. AUTHOR(S)
Daniel S. Benincasa

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
AFRL/IFEC
32 Brooks Road
Rome, NY 13441-4114

8. PERFORMING ORGANIZATION REPORT NUMBER
AFRL-IF-RS-TR-1999-162

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFRL/IFEC
32 Brooks Road
Rome, NY 13441-4114

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-1999-162

11. SUPPLEMENTARY NOTES
AFRL Project Engineer: Daniel S. Benincasa/IFEC/(315) 330-3555.

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In this research, we have developed several unique techniques used in the separation of a speech signal corrupted by another talker's speech recorded over a single channel. Historically, this has been referred to as the cocktail party problem. Our work is useful in such applications as separating the speech signals recorded onto an in-flight voice data recording box from the cockpit of an airplane, enhancing the quality of speech transmitted through a hearing aid, and in the enhancement of speech transmitted over a noisy communication channel. We have made significant contributions to the field of speaker separation. We have developed and tested an adaptive co-channel speaker separation system which can simultaneously estimate the speech of two speakers recorded onto a single channel. We have developed and tested several methods to estimate the voicing state of a co-channel speech segment. We have developed and tested a technique to estimate the fundamental frequency and pitch contour of each speaker. This technique is based on the maximum likelihood pitch estimator and harmonic magnitude suppression. Using the estimate of the fundamental frequency, we have developed a technique to estimate the harmonic parameters of overlapping voice speech segments. Finally, we have developed and tested an innovative technique to simultaneously estimate overlapping voiced speech segments using a constrained nonlinear least squared optimization algorithm. These techniques have been integrated into end-to-end speaker separation system to separate co-channel speech.

14. SUBJECT TERMS
speaker separation, co-channel interference, speech processing

15. NUMBER OF PAGES
214

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT
UL

CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES xi

ABSTRACT xix

1. INTRODUCTION 1

1.1 Historical Review 3

1.1.1 Separating Speech 4

1.1.2 Parameter Estimation 10

1.1.2.1 Voicing State Determination 11

1.1.2.2 Pitch Estimation 14

1.2 Simultaneous Adaptive Co-Channel Speaker Separation 15

2. DIGITAL SPEECH SIGNAL PROCESSING 21

2.1 System Models 22

2.1.1 Excitation and Modulation Model 22

2.1.2 Autoregressive Model 23

2.1.3 Autoregressive Moving-Average Model 24

2.1.4 Sinusoidal Model 25

2.2 Linear Predictive Coding (LPC) Analysis 26

2.2.1 Forward Prediction 26

2.2.2 Backward Prediction 28

2.2.3 Forward and Backward Prediction 29

2.3 Short Time Characteristics of Speech 29

2.3.1 Windowing Analysis 30

2.3.2 Filter Bank Analysis 34

2.3.3 Discrete Fourier Transform Spectrum and Analysis 35

2.3.4 Short Time Autocorrelation 36

2.3.5 Short Time Energy and Magnitude Measures 37

2.3.6 Cepstral Analysis 37

2.3.7 Zero-crossing Measure 38

2.3.8 Mel-Cepstrum 39

2.4 Constrained Optimization 40

2.5 Pattern Recognition 41

3. CO-CHANNEL SPEAKER SEPARATION 45

3.1 Preliminary Research 46

3.1.1 Harmonic Enhancement 47

3.1.2 Harmonic Suppression 53

3.1.3 Current Research 62

3.2 Constrained Nonlinear Least Squared Optimization 65

3.3 Adaptive Filtering (V/UV, UV/V, and UV/UV) 75

3.4 Voicing State Determination 79

3.4.1 Decision Structure 80

3.4.2 Features 82

3.4.3 Training Data 84

3.4.4 Bayesian Classifier 86

3.4.5 k-Nearest Neighbor Classifier 90

3.4.6 Parzen Window Classifier 92

3.5 Joint Pitch Estimation 93

3.5.1 Joint Pitch Estimation Techniques 94

3.5.2 Maximum Likelihood Pitch Estimation with Harmonic Suppression 101

3.6 Harmonic Selection 105

3.7 Speech Synthesis and Reconstruction 106

4. RESULTS AND ANALYSIS 108

4.1 Speech Databases 108

4.2 Parameter Estimation Analysis 113

4.2.1 Voicing State Determination Algorithm 113

4.2.2 Joint Pitch Estimation 125

4.3 Constrained Nonlinear Optimization 134

4.4 Simultaneous Adaptive Co-Channel Speaker Separation 145

5. DISCUSSION 178

5.1 Conclusions 178

5.2 Future Research 180

5.2.1 Error Measurement 180

5.2.2 Voicing State Determination 181

5.2.3 Joint Pitch Estimation 181

BIBLIOGRAPHY 182

LIST OF TABLES

Page

Table 4.1 TIMIT database speech sentences used in analysis 110

Table 4.2 Speech sentences used to train and test the VSDA. 111

Table 4.3 Speech sentence pairs used for testing male/female co-channel speech. 111

Table 4.4 Speech sentence pairs used for testing male/male co-channel speech. 112

Table 4.5 Speech sentence pairs used for testing female/female co-channel speech. 112

Table 4.6 Confusion matrix, using the Bayes classifier, of Male/Female speech mixtures with SIR = 0 dB. Overall 83.43% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses. 120

Table 4.7 Confusion matrix, using the *kNN* classifier, of Male/Female speech mixtures with SIR = 0 dB. Overall 83.62% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses. 120

Table 4.8 Confusion matrix, using the Parzen window classifier, of Male/Female speech mixtures with SIR = 0 dB. Overall 83.24% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses 121

Table 4.9 Confusion matrix, using the Bayes classifier, of Male/Male speech mixtures with SIR = 0 dB. Overall 76.49% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses. 121

Table 4.10 Confusion matrix, using the *kNN* classifier, of Male/Male speech mixtures with SIR = 0 dB. Overall 74.16% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses. 122

Table 4.11 Confusion matrix, using the Parzen window classifier, of Male/Male speech mixtures with SIR = 0 dB. Overall 77.42% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses. 122

Table 4.12 Confusion matrix, using the Bayes classifier, of Female/Female speech mixtures with SIR = 0 dB. Overall 71.99% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses. 123

Table 4.13 Confusion matrix, using the *kNN* classifier, of Female/Female speech mixtures with SIR = 0 dB. Overall 71.79% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses. 123

Table 4.14 Confusion matrix, using the Parzen window classifier, of Female/Female speech mixtures with SIR = 0 dB. Overall 72.58%

of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.................................... 124

Table 4.15 Confusion matrix, using the Bayes classifier, of Male/Female speech mixtures with SIR = -6 dB. Overall 84.21% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses... 124

Table 4.16 Percentage of pitch errors, for a male/female mixture (SIR = 0 dB), which were greater than 20%, 10%, and 5% of the normalization factor (100 Hz). The average pitch frequencies for the male and female speakers were 109.37 Hz and 186.44 Hz respectively. 132

Table 4.17 Percentage of pitch errors, for a male/male mixture (SIR = 0 dB), which were greater than 20%, 10%, and 5% of the normalization factor (100 Hz). The average pitch frequencies for the Male1 and Male2 speakers were 109.37 Hz and 157.0 Hz respectively.................. 132

Table 4.18 Percentage of pitch errors, for a female/female mixture (SIR = 0 dB), which were greater than 20%, 10%, and 5% of the normalization factor (100 Hz). The average pitch frequencies for the Male1 and Male2 speakers were 186.95 Hz and 213.82 Hz respectively.. 133

Table 4.19 Percentage of pitch errors, for a male/female mixture (SIR = -6 dB), which were greater than 20%, 10%, and 5% of the normalization factor (100 Hz). The average pitch frequencies for the male and female speakers were 109.17 Hz and 186.44 Hz respectively 133

Table 4.20 Error measurements between reconstructed signal and original signal in a male/female mixture with SIR = 0 dB. Mean represents mean error over 4 sentence combinations for male (M), female (F), and Co-channel (C) signals. 150

Table 4.21 Error measurements between reconstructed signal and original signal in a male/male mixture with SIR = 0 dB. Mean represents mean error over 4 sentence combinations for Male1 (M1), Male2 (M2), and Co-channel (C) signals. 150

Table 4.22 Error measurements between reconstructed signal and original signal in a female/female mixture with SIR = 0 dB. Mean represents mean error over 4 sentence combinations for Female1 (F1), Female2 (F2), and Co-channel (C) signals. 151

Table 4.23 Error measurements between reconstructed signal and original signal in a male/female mixture with SIR = -6 dB. Mean represents mean error over 4 sentence combinations for male (M), female (F), and Co-channel (C) signals 151

Table 4.24 Contribution and percentage of errors per voicing state, for the Male in Male/Female speech mixtures at SIR = 0 dB 159

Table 4.25 Contribution and percentage of errors per voicing state, for the Female in Male/Female speech mixtures at SIR = 0 dB. 159

Table 4.26 Contribution and percentage of errors per voicing state, for the Male1 in Male1/Male2 speech mixtures at SIR = 0 dB. 160

Table 4.27 Contribution and percentage of errors per voicing state, for the Male2 in Male1/Male2 speech mixtures at SIR = 0 dB. 160

Table 4.28 Contribution and percentage of errors per voicing state, for the Female1 in Female1/Female2 speech mixtures at SIR = 0 dB. 161

Table 4.29 Contribution and percentage of errors per voicing state, for the Female2 in Female1/Female2 speech mixtures at SIR = 0 dB. 161

Table 4.30 Confusion matrix, using the Bayes Classifier, of Male/Female co-channel speech mixture, from the RPI_COC database, SIR = 0 dB. Overall 80.21% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses............. 173

LIST OF FIGURES

Page

Figure 1.1 Segment of voiced speech. 12

Figure 1.2 Segment of unvoiced speech. 12

Figure 1.3 Pitch period defined in voiced speech segment 19

Figure 2.1 The excitation-modulation model of speech production. 22

Figure 2.2 Common window functions. 32

Figure 2.3 Comparison of the magnitude response for 512 sample length window functions: (a) rectangular, (b) Hamming, (c) Blackman, and (d) Hanning 33

Figure 2.4 Block diagram on the calculation of the mel-cepstrum coefficients. 40

Figure 2.5 Generic pattern recognition system. 42

Figure 3.1 Block diagram showing the excitation of the signal's LPC synthesis filter with the co-channel excitations signal and excitation of the co-channel LPC synthesis filter with the signal's excitation signal. 49

Figure 3.2 Example of the effects of speech enhancement. (a) Plot of output s_1 (excitation of co-channel LPCs with signal's driving function) and (b) plot of original speech signal. 51

Figure 3.3 Example of the effects of speech enhancement: (a) plot of output s_2 (excitation of original LPCs with the co-channel driving function) and (b) plot of original speech signal. 52

Figure 3.4 Solid Line Plot - spectral magnitude of co-channel speech combined in the time domain. Dashed Line Plot - sum of individual magnitude spectra. 56

Figure 3.5 Spectral Magnitude Subtraction - ideal case. 57

Figure 3.6 Results of Spectral Magnitude Subtraction when the magnitude spectrum of the interference is known and the phase of the co-channel speech is used: (a) the reconstructed speech signal, (b) the original speech signal. 59

Figure 3.7 Spectral Magnitude Subtraction with LPF - spectral magnitude subtraction is used for voiced interference and an LPF is applied when interference is unvoiced. 60

Figure 3.8 Results from Spectral Magnitude Subtraction using an LPF for unvoiced interference: (a) reconstructed speech signal, and (b) original speech signal. 61

Figure 3.9 Simultaneous Adaptive Speaker Separation 64

Figure 3.10 Representation of a Fricative, (a) time waveform and (b) magnitude spectrum. 76

Figure 3.11 Representation of a Plosive, (a) time waveform and (b) magnitude spectrum. 77

Figure 3.12 Voicing state decision tree for co-channel speech. 81

Figure 3.13 Voiced/Unvoiced segmentation of uncorrupted speech. 85

Figure 3.14 Spectrum of MCV coefficients of overlapping vocalic speech segment: (a) SIR = 0 dB, pitch frequencies measured at 99 Hz and 172 Hz, and (b) SIR = -6 dB, pitch frequencies measured at 128 Hz and 176 Hz. True pitch frequencies are 111 and 176 Hz. 96

Figure 3.15 Modified autocorrelation pitch estimation technique using nonlinear pre-processing. 97

Figure 3.16 Modified autocorrelation pitch estimation on overlapping vocalic speech at SIR = -6 dB. Pitch estimated at 113 Hz........................ 99

Figure 3.17 Maximum likelihood pitch estimation technique........................ 99

Figure 3.18 Maximum likelihood pitch estimation of overlapping vocalic speech at SIR = -6 dB. Pitch estimated at 178 Hz........................ 100

Figure 3.19 Joint pitch estimation using a maximum likelihood pitch estimator with harmonic magnitude suppression........................ 102

Figure 3.20 Maximum likelihood pitch estimation of overlapping vocalic speech at SIR = -6 dB. Pitch estimate of stronger signal measured at 178 Hz. 103

Figure 3.21 Maximum likelihood pitch estimation on residual co-channel speech after harmonic suppression of dominate harmonics. Pitch estimate of weaker signal is measured at 111 Hz........................ 104

Figure 4.1 Histogram comparing distribution between Voiced (H0) and Unvoiced (H1) feature values used in training (Level 1) on a male/female speech mixture........................ 116

Figure 4.2 Histogram displaying distribution of All Voiced (H0) and Mixed Voiced (H1) feature values used in training (Level 2) on a male/female speech mixture. 117

Figure 4.3 Histogram displaying distribution of Voiced/Unvoiced (H0) and Unvoiced/Voiced (H1) feature values used in training (Level 3) on a male/female speech mixture. 118

Figure 4.4 Comparison between pitch contour measured on uncorrupted speech (solid line) and the pitch contour measured from the co-channel speech (dotted line) for (a) male speaker, and (b) female speaker. Speech was taken from the TIMIT_COC database and mixed at SIR = 0 dB. 127

Figure 4.5 Histogram of normalized pitch errors (relative to 100 Hz) at SIR = 0 dB for male/female co-channel speech mixtures. (a) Male, (b) Female, and (c) Total for both male and female pitch errors. 128

Figure 4.6 Histogram of normalized pitch errors (relative to 100 Hz) at SIR = 0 dB for male/male co-channel speech mixtures. (a) Male1, (b) Male2, and (c) Total of both Male1 and Male2 pitch errors. 129

Figure 4.7 Histogram of normalized pitch errors (relative to 100 Hz) at SIR = 0 dB for female/female co-channel speech mixtures. (a) Female1, (b) Female2, and (c) Total of both Female1 and Female2 pitch errors 130

Figure 4.8 Histogram of normalized pitch errors (relative to 100 Hz) at SIR = -6 dB for male/female co-channel speech mixture. (a) Male, (b) Female, and (c) Total of both male and female pitch errors. 131

Figure 4.9 Reconstruction of simulated vocalic speech signal using constrained nonlinear optimization, comparing (a) reference signal 1, (b) initial signal 1, and (c) reconstructed signal 1. 137

Figure 4.10 Reconstruction of simulated vocalic speech signal using constrained nonlinear optimization, comparing (a) reference signal 2, (b) initial signal 2, and (c) reconstructed signal 2. 138

Figure 4.11 Hanning weighted speech segments, (a) male speaker, (b) female speaker, and (c) co-channel speech (male+female). 140

Figure 4.12 Male speech segment using constrained nonlinear optimization to separate co-channel speech, (a) reconstructed speech segment, (b) original speech segment. 141

Figure 4.13 Magnitude spectrum of male speech segment using constrained nonlinear optimization to separate co-channel speech, (a) reconstructed spectrum, (b) original spectrum. 142

Figure 4.14 Female speech segment using constrained nonlinear optimization to separate co-channel speech, (a) reconstructed speech segment, (b) original speech segment.. 143

Figure 4.15 Magnitude spectrum of female speech segment using constrained nonlinear optimization to separate co-channel speech, (a) reconstructed spectrum, (b) original spectrum. 144

Figure 4.16 Male speech signal, from a male/female co-channel speech mixture at SIR = 0 dB. (a) Reconstructed male speech signal and (b) original male speech signal. 148

Figure 4.17 Female speech signal, from a male/female co-channel speech mixture at SIR = 0 dB. (a) Reconstructed female speech signal and (b) original female speech signal 149

Figure 4.18 Plot of the average relative error versus increasing a priori information, for male/female speech mixtures at SIR = 0 dB. No Cheating (NC); True Voicing State (TVS); True Pitch (TP). 154

Figure 4.19 Plot of the average relative error versus increasing a priori information, for male/male speech mixtures at SIR = 0 dB. No Cheating (NC); True Voicing State (TVS); True Pitch (TP). 155

Figure 4.20 Plot of the average relative error versus increasing a priori information, for female/female speech mixtures at SIR = 0 dB. No Cheating (NC); True Voicing State (TVS); True Pitch (TP). 156

Figure 4.21 Plot of the average relative error versus increasing a priori information, for male/female speech mixtures at SIR = -6 dB. No Cheating (NC); True Voicing State (TVS); True Pitch (TP). 157

Figure 4.22 Plot of the average relative error versus SIR = 0, -6 and -12 dB for male/female speech mixtures. 164

Figure 4.23 Plot of the average relative error versus SIR = 0, -6, and -12 dB for male/male speech mixtures. 165

Figure 4.24 Plot of the average relative error versus SIR = 0, -6, and -12 dB for female/female speech mixtures. 166

Figure 4.25 Plot of the overall percent correct detect of the VSDA on male/female (M/F), male/male (M/M) and female/female (F/F) speech mixtures at SIR = 0, -6, and -12 dB.......... 167

Figure 4.26 Plot of the cumulative percent correct detect of the five voicing states of co-channel speech in the male/female speech mixtures at SIR = 0, -6, and -12 dB. 168

Figure 4.27 Plot of the cumulative percent correct detect of the five voicing states of co-channel speech in the male/male speech mixtures at SIR = 0, -6, and -12 dB. 169

Figure 4.28 Plot of the cumulative percent correct detect of the five voicing states of co-channel speech in the female/female speech mixtures at SIR = 0, -6, and -12 dB. 170

Figure 4.29 Comparison between pitch contour measured on uncorrupted speech (solid line) and from the co-channel speech (dotted line) for (a) male speaker, and (b) female speaker. Speech was taken from the RPI_COC database, SIR = 0 dB.......... 174

Figure 4.30 Male speech signal, from a male/female co-channel real speech mixture at SIR = 0 dB. (a) Reconstructed male speech signal and (b) original male speech signal 175

Figure 4.31 Female speech signal, from a male/female co-channel speech mixture at SIR = 0 dB. (a) Reconstructed female speech signal and (b) original female speech signal 176

Figure 4.32 Reconstructed speech signals from a true male/female co-channel speech signal with average SIR = 0 dB. (a) Reconstructed male speech signal and (b) reconstructed female speech signal. 177

ABSTRACT

This proposal describes several novel and unique techniques used in the separation of a speech signal corrupted by another talker's speech recorded over a single channel. Historically, this has been referred to as the cocktail party problem. Our work is useful in such applications as separating the speech signals recorded onto an in-flight voice data recording box from the cockpit of an airplane, enhancing the quality of speech transmitted through a hearing aid, and in the enhancement of speech transmitted over a noisy communication channel.

We have made significant contributions to the field of speaker separation. We have developed and tested an adaptive co-channel speaker separation system that can simultaneously estimate the speech of two speakers recorded onto a single channel. We have developed and tested several methods to estimate the voicing state of a co-channel speech segment. We have developed and tested a technique to estimate the fundamental frequency and pitch contour of each speaker. This technique is based on the maximum likelihood pitch estimator and harmonic magnitude suppression. Using the estimate of the fundamental frequency, we have developed a technique to estimate the harmonic parameters of overlapping voiced speech segments. Finally, we have developed and tested an innovative technique to simultaneously estimate overlapping voiced speech segments using a constrained nonlinear least squared optimization algorithm. These techniques have been integrated into end-to-end speaker separation system to separate co-channel speech.

Experiments have been conducted using synthetic data, simulated co-channel speech data, and real co-channel speech. The simulated co-channel speech data consisted of male/female, male/male and female/female speech mixtures. Testing of the voicing state determination, joint pitch estimation and speaker separation were conducted at signal to interference ratios (SIRs) of 0 dB, -6 dB and -12 dB. Classification of the voicing state of co-channel speech attained an 84% overall correct detection rate for male/female speech mixtures. Performance of our joint pitch estimator proved sufficient to separate two speech signals, and is experimentally evaluated. It is shown that our co-channel speaker separation system can successfully separate overlapping speech of two speakers.

1. INTRODUCTION

In a major airline crash, the in-flight voice data recording box is one of the few pieces of equipment on an aircraft that is designed to remain intact. In such a catastrophic event, this piece of equipment becomes invaluable to the investigators determining the cause of the accident. The data obtained by this recorder is a single channel audio recording of the voices, warning sounds and background noises originating from within the cockpit of the aircraft leading up to the time of the accident. The problem facing the investigators is deciphering the content of the speech signals from the recording. This is an example of a problem, historically known as the cocktail party effect, in which overlapping speech from multiple speakers is recorded over a single channel. Co-channel speaker separation is the process of extracting or separating the desired speech signal from one or more interfering speech signals.

In the past, co-channel speaker separation systems have been developed using techniques that suppress the interfering signals, enhance the desired signal or perform suppression and then enhancement to separate the desired speech signal from the co-channel signal. To date, there is no one technique that works in all situations, with most techniques performing only under limited situations and assume specific parameters of one or both speakers be given. Two parameters most widely used as *a priori* information are the voicing state and pitch contour of each speaker. This information is crucial to most, if not all, co-channel separation systems.

This thesis presents an adaptive co-channel speaker separation system that simultaneously estimates the desired and interfering speech signals. Our system first estimates the voicing state and pitch contour of each speaker. Voicing state classification is based on whether the co-channel signal contains speech which is produced from voiced sounds, unvoiced sounds, silence or some combination of these states. Therefore, our determination algorithm decides when a given frame of co-channel speech contains voiced speech from one or both speakers, a mixture of voiced and unvoiced speech, unvoiced speech from one or both speakers, or silence. An estimate is made of the fundamental frequency of both speakers for all co-channel speech segments, including those which have been determined to contain two voiced overlapping speech segments. The fundamental frequency estimates are used to provide initial values of the sinusoidal speech model parameters. These parameters include the phase, amplitude and center frequency of the dominant spectral harmonics for each speaker. Final values of these parameters are obtained using a constrained nonlinear least squared optimization algorithm which is used to minimize the squared error between the sum of our estimated speech segments and the true co-channel speech segment. Co-channel speech segments that contain a mixture of voiced and unvoiced speech sounds or those segments in which one or both of the speakers are silent are separated using conventional filtering techniques. An overlap and add technique is used to reconstruct the original speech signals into natural sounding speech.

In this research we have made significant contributions to the area of speaker separation. We have developed and tested a simultaneous adaptive speaker separation system that separates speech from two speakers recorded over a single channel. We have developed and tested a voicing state determination algorithm to estimate the voicing state of each speaker present. We have developed and tested a joint pitch estimation algorithm to estimate the pitch contours for two speakers. We have developed a technique that uses these pitch frequencies to estimate the dominant spectral harmonics for each speaker. Finally, we have developed and tested a technique to separate overlapping voiced speech from two speakers. We have conducted experiments using simulated and real speech signals in varying signal to interference ratio (SIR) environments to evaluate the performance of our algorithms and our co-channel speaker separation system.

1.1 Historical Review

Historically, speaker separation systems have been focused around suppression of the interfering speech signal or enhancement of the desired speech signal. Very few have concentrated on the estimation of both signals simultaneously. Regardless, all systems developed thus far require some form of *a priori* information on each of the speakers present. One such *a priori* parameter that has been widely used is pitch. Here, we will provide a brief historical review of the work in speaker separation, voicing determination and pitch estimation to obtain a more in-depth understanding as to the level of accomplishments in each area. This will also provide an opportunity to include some of the work performed in pitch estimation and voicing state determination of uncorrupted speech that has impacted our research.

1.1.1 Separating Speech

Research that resulted in significant improvement to the enhancement and intelligibility of a speaker in the presence of another speaker or in the presence of noise dates back to the 1970's with the work of Mitchell [1] and Parsons [2]. The work by Mitchell attempted to solve the cocktail party effect using a class of nonlinear processes for the output of an array of two (or more) microphones. It was shown that nonlinear processing proved to be more effective at suppressing unwanted sources when the unwanted source is near one of the microphones. This is the binaural problem, which will not be addressed in this work. Here, we will only be concerned with the case where speech has been recorded using a single microphone.

Parsons was one of the first researchers to apply harmonic selection to suppress the effects of interfering speech *without* the benefit of binaural information. Parsons required that both voices be periodic, restricting the separation process to sounds which are vowels or vowel-like. His work dissected the Fourier transform of the signal into components belonging to each talker. Pitch tracking was used as a means of providing continuity of talker identification from segment to segment. His later work attempted to improve his initial system ([3],[4]). Improvements included extending the capability of his system to perform in a natural speech environment, to improve the intelligibility of the restored speech and to handle non-vocalic speech sounds. Parsons indicated that errors in pitch estimation resulted in the unintelligibility of restored speech using harmonic selection. He continued his work attempting to improve the detection of pitch as a way to separate speech but in a paper published in 1979 [5], he concluded his research with no significant improvements over the original system.

Shields [6] and Frazier [7] investigated separating speech signals by means of a variable comb filter that passed only the desired talkers pitch harmonics. In both cases the pitch was assumed to be a known parameter. Everton [8] used pitch and formant information to drive a vocoder. Here, as with the previous work, the pitch was assumed known while formant information was derived from inspection of the combined speech spectrum. This was accomplished by sampling the log-magnitude spectrum of the co-channel speech signal at harmonics of the fundamental frequency. These samples were then used to construct a rough curve, which was then lowpass filtered to reconstruct an outline of the formant structure of a single speaker.

Dick [9] performed some experiments with an adaptive comb filter tuned to the pitch harmonics of one particular speaker along with the use of the complex correlation. However, a major problem still remained; how to identify which speaker is making which sounds solely from a co-channel speech signal.

Hanson and Wong ([10],[11],[12]) attempted to improve the intelligibility of overlapping speech using Harmonic Magnitude Suppression (HMS). This method has two distinct components, an interfering speech estimator that samples the short-term magnitude spectrum at the harmonics of the interfering fundamental frequency and an interference remover, which is based on spectral magnitude subtraction. Extensive testing indicated intelligibility improvement for negative dB signal to interference ratio

(SIR) cases. However, their technique still required an *a priori* estimate of the pitch of the interfering speaker and a determination of the voicing state of each speaker.

Lee and Childers [13] proposed a method based on a two-stage approach. The first stage produced an initial estimate of the speech spectrum for each talker. The second stage used this initial estimate to spectrally tailor the spectrum of each talker, taking into account the autocorrelation function of the known co-channel composite speech signal. This work only showed slight improvement over that of Hanson and Wong and required the pitch harmonics of the speakers to be well separated for best results.

Wientrab [14] used pitch as an input to a Markov Model to identify the number of people talking within a given time interval and the type of sounds being produced by each person present. He classified the sounds using one of three steady-state labels (silence, periodic, or non-periodic). Performance of the Markov Model decreased significantly when pitch tracks extracted from the co-channel speech signal were used as opposed to *a priori* pitch tracks taken from single speech signals before being corrupted.

Alexander [15] used an adaptive Least Mean Squares (LMS) algorithm to separate target speaker produced information from non-target speaker produced noise based on the difference in power levels associated with the two phenomena.

Naylor and Boll [16,17] continued the work of Hanson and Wong by estimating certain model parameters solely from the co-channel speech and by improving the technique for suppressing non-voiced interference. They investigated techniques for pitch tracking and magnitude suppression. The pitch estimate of the louder talker was measured from the co-channel signal. Their conclusion on the performance of the HMS technique developed by Hanson and Wong was limited by the degree to which the actual speech deviates from the ideal model and by the degree to which speaker specific parameters are recoverable from the co-channel signal.

Kopec and Bush [18] investigated an LPC-based spectral similarity measure to perform speech recognition in a co-channel speech environment. It was assumed that the interfering speaker would introduce extra poles (corresponding to formants) into the LPC spectrum of co-channel signal. An attempt was made to determine which poles were associated with the desired speaker and which poles resulted from the interfering speaker. A reference spectrum was used to hypothesize whether a pole was part of the desired speech or from the interfering speech. The reference spectrum was very similar to the spectrum of the desired speaker. In isolated word recognition experiments, the authors demonstrated that error rates could be reduced by up to 70% at low SIRs.

Rogers et al [19] developed an automated algorithm for multiple speech separation based upon a variable frame-size orthogonal transform and a spectral matching technique. A multi-step pitch detection scheme, which relied on the traditional autocorrelation function was also used. Improvements to this system later included a neural network to predict the number of speakers present and the voicing state of each speaker for each frame of co-channel speech data. This work relied on an energy threshold to determine the number of talkers present in the co-channel speech. These systems still required accurate *a priori* information as to the voicing state of each speaker present and the pitch of each speaker.

Varga and Moore [20] used the Hidden Markov Model (HMM) to separate speech and noise, however tests were only conducted on speech signals corrupted by periodic noise, not with another speech signal.

Gish [21] investigated a clustering technique to separate co-channel speech signals. This method requires no *a priori* information and uses a distance measure between speech segments based on the likelihood ratio of speech segments using multivariate Gaussian assumptions.

Gu and van Bokhoven [22] developed a new approach to speaker separation using frequency bin nonlinear adaptive filtering along with a robust multi-pitch estimation routine which uses HMMs to simultaneously estimate the pitch of multiple speakers from the co-channel signal. Limited testing, using synthetic speech of two speakers with fixed but different pitches, showed improvement on attenuating most of the interfering speech signal.

Quatieri and Danisewicz [23] developed a technique which combined modeling voiced speech as a sum of sinusoids with time-varying amplitudes along with a linear least mean-squared error estimation procedure to separate overlapping spectral harmonics. Here, they assumed that the harmonic frequencies were exact integer multiples of the fundamental frequency and only processed voiced speech. One major conclusion of this work is that further development of a pitch estimation algorithm, capable of handling summed waveforms of vastly different intensity levels, is crucial to speech separation and enhancement.

Naylor and Porter ([24],[25]) investigated a unique technique to estimate the pitch of multiple speakers in a co-channel speech environment based on a Modified Covariance (MCV) spectrum estimator. They also investigated a linear estimation technique for resolving the complex spectrum of the co-channel signal into individual speakers' components. Testing was limited to using simulated data.

Zissman ([26],[27]) concentrated on suppression of a jamming speech signal at target to jammer ratios ranging from -3 to -15 dB. Results indicated a 10 to 20 dB jammer attenuation, providing improved target intelligibility. He also addressed the problem of speaker activity detection, the labeling of co-channel speech as target-only, jammer-only or two-speaker (target plus jammer) speech using speaker identification techniques.

The work by Savic et al. [28] developed a speaker separation system based on a Maximum Likelihood Deconvolution (MLD) that would simultaneously estimates the excitation signal of multiple speakers based on an estimate of the vocal tract filters of each speaker. The proposed MLD technique, required no *a priori* information, but was never demonstrated under this condition. Successful demonstration of this technique using parameters extracted from the speech signal *a priori* were conducted. They were unable to obtain the necessary speaker specific parameters strictly from the co-channel signal.

Finally, Morgan et al. [29] proposed a co-channel speaker separation system using a Harmonic Enhancement and Suppression (HES) technique. This system relied on an accurate estimate of the pitch of each speaker present. A maximum likelihood pitch detector [30] was used to estimate the pitch of the stronger speaker. Morgan et. al. also proposed a maximum likelihood speaker assignment (MLSA) algorithm to label the recovered stronger and weaker signals as coming from either the target or the interfering speaker. Recent published results showed limited success [31].

New work in the area of co-channel speaker separation is being conducted at Rutgers University using energy separation techniques to estimate the amplitude and instantaneous frequency of each speech signal. Using a nonlinear differential operator they are looking to detect modulations in AM-FM signals by estimating the product of their time-varying amplitude and frequency [32].

1.1.2 Parameter Estimation

Parameter estimation is one of the more crucial and difficult aspects of speaker separation. This is an area in which most techniques assume information is known prior to separation. In order to have a realistic and robust system, all parameters must be estimated solely from the co-channel speech signal. The parameters required for this research as well as for most other systems, include the voicing state of the speakers present, the pitch period of each speaker during phonation and the phase, amplitude and center frequency of the spectral harmonics for each speaker.

1.1.2.1 Voicing State Determination

Voicing state determination is a process by which a segment of speech is classified as voiced speech, unvoiced speech or silence. A voiced sound is one in which phonation, the oscillation of the vocal cords, is present [33]. An example of the time waveform of a voiced speech segment is shown in Figure 1.1. Characteristics of a voiced speech include a periodic nature of the time waveform with relatively high energy. An unvoiced sound can be classified as a sound that is not voiced. An example of an unvoiced segment of speech is shown in Figure 1.2. Unvoiced speech is noise-like in nature with relatively low energy. Silence is classified as the absence of speech altogether.

For an uncorrupted speech signal, typical voicing state determination algorithms can be grouped into three different categories, simple threshold analysis algorithms, complex algorithms based on pattern recognition techniques and integrated algorithms which make both a voicing determination and pitch estimation simultaneously. Very few attempts have been made at voicing state determination on a co-channel speech signal. A pattern recognition technique was applied to co-channel speech to determine the number of speakers present in the co-channel speech with testing performed on speech segments from known subjects used in the training [26]. Techniques that attempt to estimate the pitch and voicing state simultaneously, it is incorrect to rely on the presence of pitch as a means of making a voicing determination. Pitch can only exist when the signal is voiced. However, it is incorrect to assume that a segment of speech is unvoiced simply because a pitch period does not exists or is not measurable.

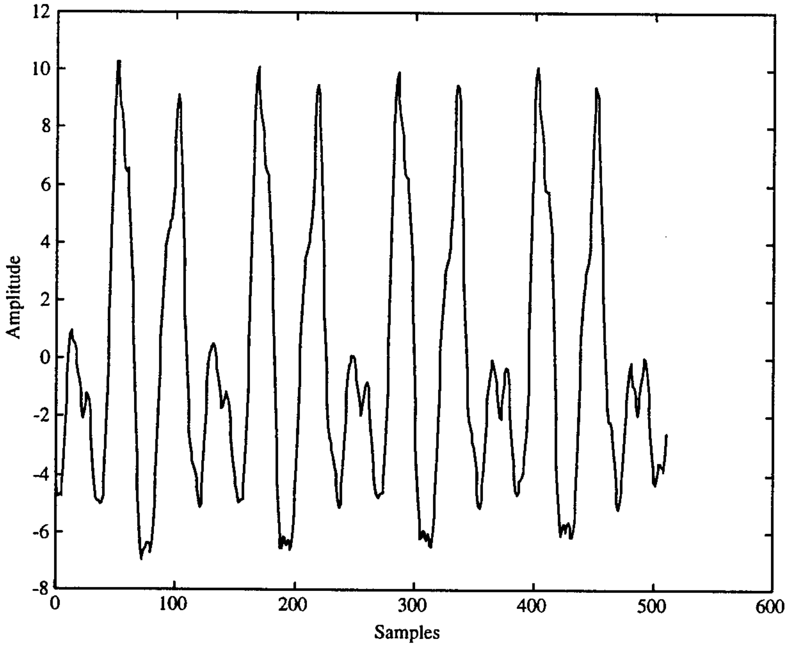


Figure 1.1: Segment of voiced speech.

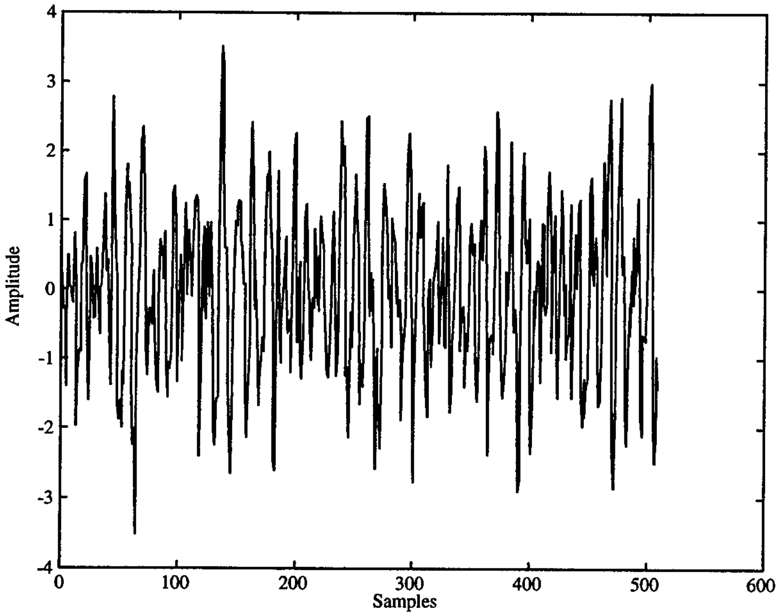


Figure 1.2: Segment of unvoiced speech.

Several articles have been written on the voiced classification of uncorrupted speech waveforms dating back to the 1960's in which the pitch period was used as a means of identifying whether the segment of speech was voiced. This technique was shown to be unreliable. Rabiner and Atal [34] used a classical statistical pattern recognition approach that made a three-class decision of telephone speech. Here the parameter set was narrowed down to five different measurements on the speech signal; zero-crossing rate, speech energy, the correlation between the adjacent speech samples, the first predictor coefficient from a 12-pole linear predictive coding (LPC) analysis and the total energy in the prediction residual. Later, Rabiner and Sambur [35] performed classification by nonlinearly combining an LPC distance measure and an energy distance measure to discriminate between the three classes. This technique demonstrated a low error rate for segments of speech confined to a single class. Siegel and Bessey [36] used a Bayesian classifier in a binary decision tree structure in which the speech segment was first classified as predominately voiced or unvoiced. The segment was then tested to determine if the excitation of the segment contained a mixture of voiced and unvoiced speech. The feature set included 14 distinct features similar to the ones used by Rabiner and Atal.

In most co-channel speaker separation systems, researchers process only those frames in which the desired speaker is voiced. Zissman applied speaker identification techniques to determine the presence or absence of a particular speaker in a co-channel speech signal [26]. The technique was tested using two separate classifiers, a vector-quantizing classifier and a modified Gaussian classifier. A codebook was created by training a feature vector, 20 mel-frequency weighted cepstral coefficients, on the speech from the desired speaker, speech corrupted by an interfering speaker and on the interfering speech alone for a variety of SIRs. Both classifiers worked relatively well using supervised and unsupervised learning.

Morgan et al. [31] attempted to classify each segment of the co-channel signal as either voiced or silence/unvoiced. Classification was performed by extracting five features from frames of speech and analyzing them using a multivariate Gaussian classifier. The classifier generates a binary voicing decision related to the presence or absence of voiced speech. This system only identifies the presence or absence of voiced speech, it does not identify the voicing state of each speaker.

1.1.2.2 Pitch Estimation

The pitch period or fundamental frequency of speech corrupted by another speech signal, based solely on the co-channel waveform, has been assumed known by most researchers. Pitch is a critical parameter from the standpoint that most of the systems attempt to process on the harmonics of the desired or interfering speaker. These harmonics are located near integer multiples of the pitch frequency. This requires an accurate estimate of the pitch frequency of one or both speakers.

Woodsman et al [37] developed a two-speaker pitch estimation algorithm in which an accurate estimate of the pitch of the interfering speaker was obtained for SIRs ranging from 0 to -9 dB. However, independent testing by Zissman using synthetic vowels could not confirm his results [26].

Naylor and Porter [24] implemented a Modified Covariance (MCV) spectral estimator in which it was shown that the harmonics of both speakers were clearly evident in the MCV spectra of a voiced/voiced co-channel speech signal. A clustering algorithm was developed to extract the pitch estimates from the MCV spectra. Their preliminary testing demonstrated promising results.

1.2 Simultaneous Adaptive Co-Channel Speaker Separation

Our speaker separation system accurately performs five major functions in order to separate co-channel speech. First we classify the voicing state of each frame of co-channel speech. Second, we measure the pitch frequency within those frames in which one or both of the speakers are voiced. Third, we estimate the spectral harmonics for each speaker. Fourth, the co-channel speech segments are processed based on the voicing state of each speaker. Last, the speech segments are reconstructed in such a manner to form naturally sounding and intelligible speech.

In this thesis, we have developed a voicing state determination algorithm. This algorithm classifies co-channel speech (on a frame-by-frame basis) as belonging to one of five classes. These classes include silence (both speakers are silent), unvoiced/unvoiced (both speakers are producing unvoiced sounds), voiced/unvoiced (desired speaker is voiced, interfering speaker in unvoiced), unvoiced/voiced (desired speaker is unvoiced and interfering speaker is voiced) and voiced/voiced (both speakers are voiced).

Voicing is defined as the presence or absence of phonation. Voiced speech is present during phonation, which occurs when the vocal cords are excited by a series of impulses, or bursts of air which resonate through the vocal tract. Voiced speech typically has high energy and is somewhat periodic and stationary in appearance. Unvoiced speech is the absence of phonation. This speech is typically represented by turbulent airflow past some constriction in the vocal tract. Unvoiced speech in appearance is typically of low energy and noise-like. Refer back to Figure 1.1 and Figure 1.2. The voiced/unvoiced determination problem is deciding whether the vocal cords are vibrating within a segment of speech. We can group sounds into three basic classes:

1. Vocalic sounds are produced by exciting the vocal tract with quasi-periodic pulses of airflow caused by the opening and closing of the glottis.
2. Fricative sounds are produced by forming a constriction somewhere in the vocal tract and forcing air through the constriction so that turbulence is created, thereby producing a noise like sound.
3. Plosive sounds are produced by completely closing off the vocal tract, building up pressure behind the closure, and then abruptly releasing it.

In this research, we are concerned with detecting the presence or absence of phonation of two speech signals that have been recorded onto a single channel. We have assumed a scenario in which the interfering speaker has an *average* signal strength that is equal to or stronger than the desired speaker. These are signals in which the overall average signal to interference ratio (SIR) is less than or equal to 0 dB. We define the SIR, in decibels, as 20 times the log of the ratio of the two-norm of desired speech signal to the two-norm of the interfering speech signal

$$SIR = 20 * log\left(\frac{\left(\sqrt{\sum_{j=1}^{N} s_D(j)^2}\right)}{\left(\sqrt{\sum_{j=1}^{N} s_I(j)^2}\right)}\right) \tag{1.1}$$

This however may create a situation in which the unvoiced speech of the interfering speaker is at the same or at a higher energy level than the voiced speech of the desired speaker. Or during the onset or offset of voicing of the undesired speaker, an overlapping voiced sound from the desired speaker may be stronger than the undesired voiced sound. Herein lies the difficulty of identifying the voicing state of co-channel speech.

The pitch frequency or fundamental frequency is the reciprocal of the vocal cord vibration period (fundamental period) due to the opening and closing of the vocal cords. Pitch can also be thought of as the repetition rate of the pulses in an excitation signal that produced voiced speech. See Figure 1.3. Males and females have different pitch ranges to which he or she is physically constrained. For males the possible pitch range is between 50 and 250 Hz while for females the range is somewhat higher, 120 to 500 Hz. There have been many techniques developed to estimate the pitch of a single speaker [41] but few have been developed to estimate the pitch of multiple speakers recorded over a single channel.

Our technique of joint pitch estimation is to first apply the maximum likelihood pitch estimation technique to a windowed co-channel speech segment. This will provide a pitch estimate of the stronger, voiced speaker. We then use this pitch estimate to suppress the dominate spectral harmonics of the stronger speaker. For co-channel speech that contains overlapping voiced speech, we process the residual signal resulting from the harmonic suppression to calculate the pitch of the weaker speaker. In situations when the two speakers are producing voiced speech, an estimate of the average pitch for each speaker is used to assign our pitch measurements to a particular speaker, otherwise the results from the voicing state determination algorithm assigns the pitch measurement.

When the interfering signal and a desired signal are both speech signals, it is not always possible to apply conventional filtering techniques to separate the desired signal from the interfering signal. Long-term spectral characteristics of speech signals are typically not similar. Therefore, we must concentrate on exploiting those short-term spectral and temporal characteristics that allow us to separate the desired signal from the interfering signal. For speech synthesis it has been stated that "aural perception depends only on the amplitude spectrum of a sound and is independent of the phase angles of various frequency components contained in the spectrum" [38]. This is only true for reconstruction and synthesis. Here, we are dealing with the separation of overlapping speech signals. The phase of both signals becomes a crucial element in spectral estimation.

Several scenarios can be present when a desired speech signal is corrupted by an interfering speech signal. For the case of the interfering signal having more signal energy than the desired signal, both the voiced and unvoiced interfering signal may dominate the co-channel signal independent of the voicing state of the desired speech signal.

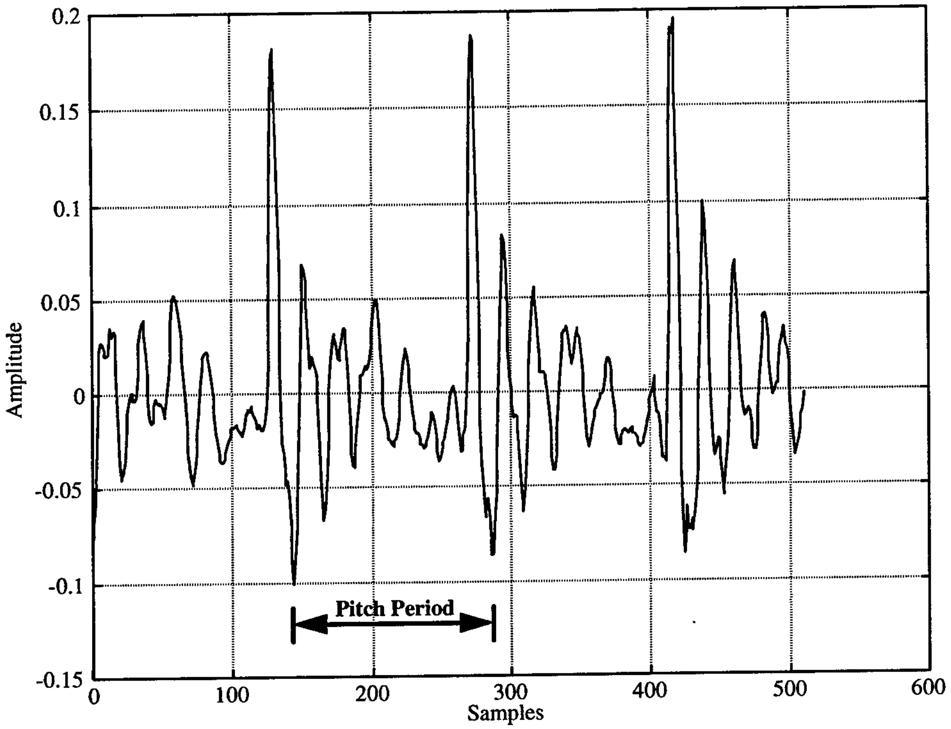


Figure 1.3: Pitch period defined in voiced speech segment.

An interfering signal that is unvoiced will typically be a signal that has the major portion of its spectral energy concentrated at frequencies higher than a voiced signal. Removing the effects of the interfering signal can just be a matter of selecting the appropriate cut-off frequency of a low-pass filter to eliminate the major portion of the interfering spectral characteristics. When both signals are unvoiced, intelligibility of a speech signal will not significantly be effected if the unvoiced speech segment is replaced with a noise-like signal.

To remove the effects of an interfering voiced signal, we need to identify the center frequencies, magnitude and phase of the spectral harmonics of that signal. When the desired signal is also voiced, it may occur that the interfering spectral harmonics overlap the desired spectral harmonics. Any form of suppression could also eliminate the harmonics of the desired signal. Therefore, one must simultaneously estimate the waveform of both signals to preserve their true spectral characteristics. We have developed a constrained nonlinear least squared optimization algorithm to simultaneously estimate the spectral harmonics of both voiced speech segments.

Reconstruction of the desired and interfering speech signals is accomplished using an overlap and add technique. With the use of a Hanning weighted window, the reconstructed speech segments are concatenated together to form intelligible and natural sounding speech.

2. DIGITAL SPEECH SIGNAL PROCESSING

This chapter introduces the digital signal processing tools that are used in this research. These tools rely on the fact that although speech is a non-stationary process, the properties of a speech signal will change relatively slowly with time. Given this assumption, it becomes beneficial to break the data into short segments, called frames, which exhibit quasi-stationary properties. Each frame of data is then processed separately. Speech segments that are too long cause our assumption of stationarity to become invalid. However, segments that are too short in duration do not provide an accurate estimate of the spectral features. Previous work has shown that a time interval of 30 msec. is adequate to insure stationarity while still providing adequate spectral resolution.

This chapter is broken into five sections. The first section deals with the system models used for speech production. The second section discusses Linear Predictive Coding (LPC) analysis which has become a crucial tool in the analysis of speech waveforms. The third section describes the short time characteristics of speech. The fourth section discusses constrained optimization and the last section discusses pattern recognition.

2.1 System Models

2.1.1 Excitation and Modulation Model

One of the most widely used models for speech production is the excitation and modulation model. As shown in Figure 2.1, the organs of the vocal tract, represented as a modulating filter are excited by a driving or excitation function. The vocal tract can be modeled as an acoustical tube. The driving function typically represents one of several different types of glottal excitation, including phonation, whispering, frication, compression, and vibration. The excitation acts as a carrier signal with the vocal tract filter modulating the acoustical information onto the excitation signal.

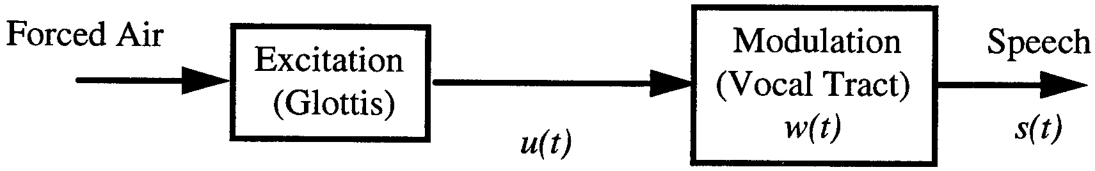


Figure 2.1: The excitation-modulation model of speech production.

In the time domain, this model is represented as

$$s(t) = u(t) \otimes w(t) \tag{2.1}$$

where *u*(*t*) is the excitation signal, *w*(*t*) is the modulation function and *s*(*t*) is the speech signal which is a convolution of the excitation signal with the modulation signal. In the frequency domain (2.1) becomes

$$S(\omega) = U(\omega)W(\omega) \tag{2.2}$$

where now $U(\omega)$, $W(\omega)$ and $S(\omega)$ represent the Fourier transform of the excitation signal, the modulation function and the speech signal respectively.

The excitation signal or glottal waveform is typically represented as either a train of impulses, or as a white noise driving source, depending on the type of sound being produced. If we lump the different types of speech into one of three classes, voiced, unvoiced or silence, then for voiced speech the excitation signal would be a train of impulses, for unvoiced speech it would consist of white noise and for silence the excitation signal would be zero.

Acoustically, we can think of this modulation as a means of filtering. The vocal tract is like any other acoustical tube in which there are natural frequencies which are a function of its shape. In speech, these natural frequencies are called formants which account for the primary way of modulating the excitation signal in the production of all the vowels and some of the consonants.

2.1.2 Autoregressive Model

Another method used to model speech production is to model the vocal tract as an autoregressive (AR) process or all-pole filter. The all-pole filter contributes to the short-time spectral envelope of a speech spectrum. In general the short-time envelope of the speech signal includes contributions from both poles and zeros. However, if we increase the number of coefficients associated with this filter, an all-pole model can approximate the effect of both the poles and zeros on the spectrum. An AR filter can be written as

$$\hat{s}[n] = -\sum_{i=1}^{p} a_i s[n-i] \tag{2.3}$$

Here the dependent variable $\hat{s}[n]$ is written as a linear combination of the independent variables $s[n\text{-}1]$ through $s[n\text{-}p]$, where p is the order of the filter. The coefficients a_i can be obtained as solutions to a set of p linear equations.

Using this model, the transfer function of the vocal tract filter can then be represented in the complex z-domain as

$$W(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p}} \tag{2.4}$$

2.1.3 Autoregressive Moving-Average Model

The all-pole model of the vocal tract filter is a method widely used in speech processing. When the order of the filter is large, it can approximate the contributions of the zeros. However, the problem of modeling the vocal tract with a pole-zero model is a classical one in nonlinear estimation theory. Most methods are based on an iterative pre-filtering scheme which is just a linearization of the same nonlinear problem. The iterative pre-filtering method is a means of estimating the poles and zeros of the transfer function, representing the vocal tract, simultaneously. Other methods obtain the parameters of a pole predictor and those of the zero predictor separately [41].

The pole-zero modeling has been shown to be effective in representing nasal sounds and some consonants, as well as an effective method in spectral estimation of noisy speech. Here the speech signal is represented as

$$\hat{s}[n] = -\sum_{i=1}^{p} a_i s[n-i] + \sum_{i=0}^{q} b_i u[n-i] \tag{2.5}$$

where the coefficients of the pole predictor are represented by the a_i coefficients and the b_i coefficients represents those of the zero predictor. The transfer function for this model of the vocal tract filter is given, in the complex z-domain as

$$W(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_q z^{-q}}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p}} \tag{2.6}$$

2.1.4 Sinusoidal Model

Based on the assumption above, in which speech is modeled as a slowly-varying vocal tract filter with a quasi-periodic train of impulses or white noise driving source, a voiced speech waveform can be represented by a sum of sine waves with time-varying amplitude, frequency and phase terms. With this assumption, our speech signal, $s[n]$ can be written as

$$s[n] = \sum_{k=1}^{M} a_k[n] cos[\theta_k[n]] \tag{2.7}$$

where the time-varying amplitude and phase terms are denoted by $a_k[n]$ and $\theta_k[n]$, respectively. The time varying frequency of each sine wave is given by the derivative of the phase, denoted by $\omega_k[n] = \theta'_k[n]$.

2.2 Linear Predictive Coding (LPC) Analysis

Linear prediction is a method by which a signal can be represented as a linear combination of it past or future values along with the current value of the input. When only the past and current values of the input are considered the analysis is said to be a forward linear prediction. An analysis which considers only the future and current values of the signal is said to be a backward linear prediction. An analysis can also consider the past, future, and current values simultaneously to form a more accurate representation. This is referred to as a forward and backward prediction.

2.2.1 Forward Prediction

The forward linear prediction estimate is given by

$$\hat{s}^f[n] = -\sum_{i=1}^{p} a_i^f s[n-i] \tag{2.8}$$

where the a_i^f represent the forward linear prediction coefficients. The prediction is forward in the sense that the estimate at time index n is based on the p samples indexed earlier in time. The linear prediction error or residual signal is just the difference between the estimate and the true signal

$$e^f[n] = s[n] - \hat{s}^f[n] \tag{2.9}$$

The solution for the optimal LPC coefficients can be calculated by minimizing the mean squared prediction error $\overline{e^{f^2}[n]}$ over the given time interval. The coefficients are obtained by setting the partial derivative of $\overline{e^{f^2}[n]}$ with respect to a_i^f equal to zero. This leads to a set of linear equations of the form

$$\begin{bmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{p1} & \cdots & c_{pp} \end{bmatrix} \begin{bmatrix} a_1^f \\ \vdots \\ a_p^f \end{bmatrix} = -\begin{bmatrix} c_{01} \\ \vdots \\ c_{0p} \end{bmatrix} \tag{2.10}$$

where $c_{ij} = \sum_i s[n-i]s[n-j]$. Since speech is quasi-stationary over short intervals of time, when we consider a different time interval we obtain a different set of coefficients.

Several techniques have been developed to solve this set of linear equations. One widely used technique to obtain the solution to (2.10) is the maximum entropy method, or the autocorrelation method [41]. The advantage of this method is that the filter formed from these coefficients is a stable filter.

When only a short segment of the data is available, the c_{ij} coefficients can be written as

$$c_{ij} = \sum_{n=0}^{N-1-|i-j|} s[n]s[n+i-j] = R(|i-j|) \tag{2.11}$$

Substituting (2.11) into equation (2.10) and solving gives the autocorrelation solution of the predictor. This solution minimizes the variance of the prediction error over all time considered. This method is also called the data windowing method since it only requires the data within a particular segment or frame.

There are other methods which will solve this linear set of equations, such as the covariance method which windows the prediction error $e^f[n]$. In this case, $N+p$ samples of the signal are needed. However, it will not guarantee a stable linear prediction filter. In speech analysis we are concerned with using the linear prediction filter to resynthesize the speech signal, therefore filter stability becomes a requirement that tends to favor the autocorrelation method.

2.2.2 Backward Prediction

The backward linear prediction estimate is given by

$$\hat{s}^b[n] = -\sum_{i=1}^{p} a_i^b s[n+i] \tag{2.12}$$

where the prediction is backward in the sense that the estimate at time index n is based on p samples indexed later in time. The backward linear prediction error or residual signal can be written as

$$e^b[n] = s[n-p] - \hat{s}^b[n-p] \tag{2.13}$$

where the error index is written with respect to n rather than n-p. This allows the forward and backward prediction errors to be functions of the same set of data samples that would be present in the linear prediction filter.

Techniques for solving the coefficients of the backward linear prediction estimate are similar to those for the forward linear prediction estimate. However, given a finite segment of data, the backward linear prediction coefficients determined by the autocorrelation method are not, in general, identical to the forward prediction coefficients.

2.2.3 Forward and Backward Prediction

Marple [41] provides us with the modified covariance (MCV) algorithm used to minimize the average linear prediction squared error in the forward and backward direction over a given segment of data. This results in a more accurate estimate of the autoregressive coefficients. The modified covariance method, however, does not guarantee a stable linear prediction filter. Therefore certain considerations must be made when it is used for purposes other than spectral estimation. Later, in the next chapter, we will see how this method can be applied to estimating the pitch frequency of multiple speakers from a co-channel signal and as a feature vector used to identify the voicing state of a speech segment.

2.3 Short Time Characteristics of Speech

As stated earlier, speech is a nonstationary process. However, if we segment a speech signal into windowed frames whose length is relatively short, the properties of the signal change only slightly during that interval of time. When the window length is too long, the signal properties may change significantly over the time interval. If the window length is too short, resolution of narrowband components may be sacrificed. This section presents the benefits gained by segmenting the speech into frames to create data segments in which features and statistics become quasi-stationary over a given time interval.

2.3.1 Windowing Analysis

Segmentation of a sampled continuous function, such as a speech signal is accomplished by a method known as windowing. Windowing is a process of multiplying a signal by a window function $w[n]$ of finite duration. This will truncate a speech signal into a finite duration segment called a frame. By delaying or advancing $w[n]$ we can examine any part of our speech signal.

Two points must be considered when choosing a window function. First we must look at the tradeoff of bandwidth versus leakage suppression. Windowing tends to broaden impulses in the theoretical Fourier representation. Thus, exact frequencies are less sharply defined. Secondly, we must choose a windowing function which will be compatible with our overlap and add reconstruction processing used at the output of our system. Refer to Figure 2.2 and Figure 2.3 for the time and spectral representation of the several common window functions.

A windowing function that works well for one application will not necessarily work well for another. The simplest window function is the rectangular window given by

$$w[n] = \begin{cases} 1 & |n| \leq N/2 \\ 0 & otherwise \end{cases} \tag{2.14}$$

where N is the frame length, or number of samples within that window. As the number of samples N increases, the width of the main lobe will decrease. However, the rectangular window is not a good choice for the overlap-and-add reconstruction process. It will introduce erroneous click sounds into a reconstructed speech signal.

Another type of window function widely used in speech processing is the Hamming window. The Hamming window is defined as

$$w[n] = \begin{cases} .54 - 0.46\,cos(4\pi n/N) & |n| \leq N/2 \\ 0 & otherwise \end{cases} \tag{2.15}$$

which is similar in shape to the raised cosine pulse. As can be seen in Figure 2.3, the main lobe of a Hamming window spectrum is wider than for the rectangular pulse, but has less energy in the side lobes. However a drawback to the Hamming window is that the taper does not go to zero at either end of the window.

The function which represents the Blackman window is defined as

$$w[n] = \begin{cases} 0.42 - 0.5\,cos(4\pi n/N) + 0.08\,cos(8\pi n/N) & |n| \leq N/2 \\ 0 & otherwise \end{cases} \tag{2.16}$$

The Blackman window is similar in shape to the Hamming window with a slightly sharper taper at either end of the window. Spectrally this will cause a slightly larger main lobe than the Hamming but with sidelobes which are distinctly lower. See Figure 2.3.

The Hanning window is represented as

$$w[n] = \begin{cases} .5 - 0.5\,cos(4\pi n/N) & |n| \leq N/2 \\ 0 & otherwise \end{cases} \tag{2.17}$$

It has a relatively narrow main lobe (although it is slightly wider than the Hamming) and lower energy in the side lobes. In the time domain, the tails of the Hanning window go to zero which is an advantage, when applied to the overlap-and-add reconstruction process. It will be the window of choice in this research.

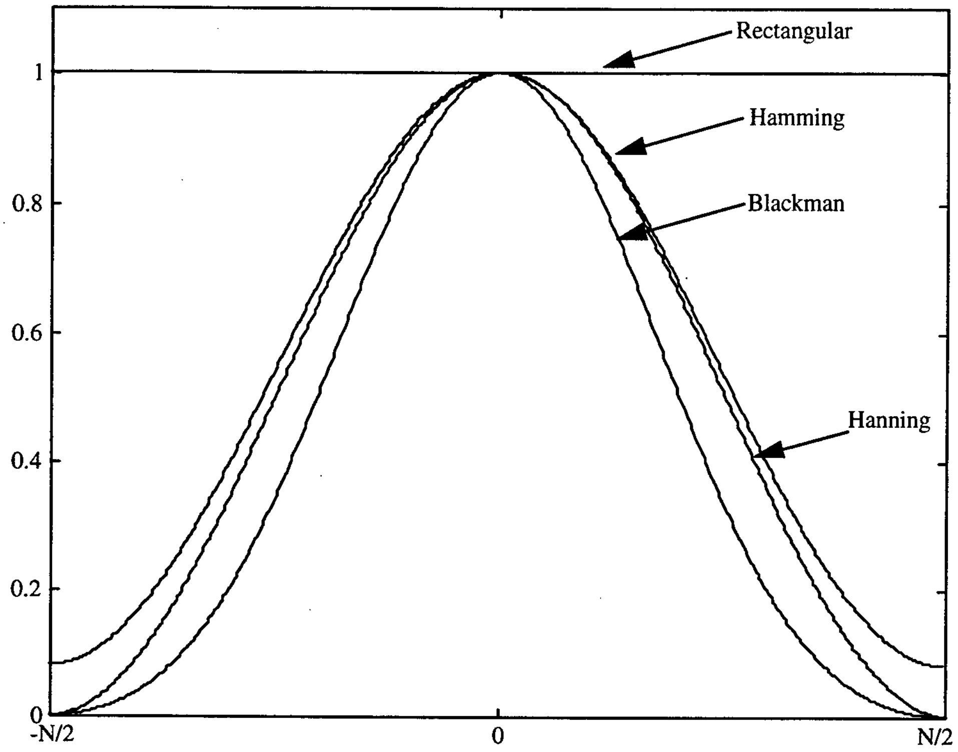
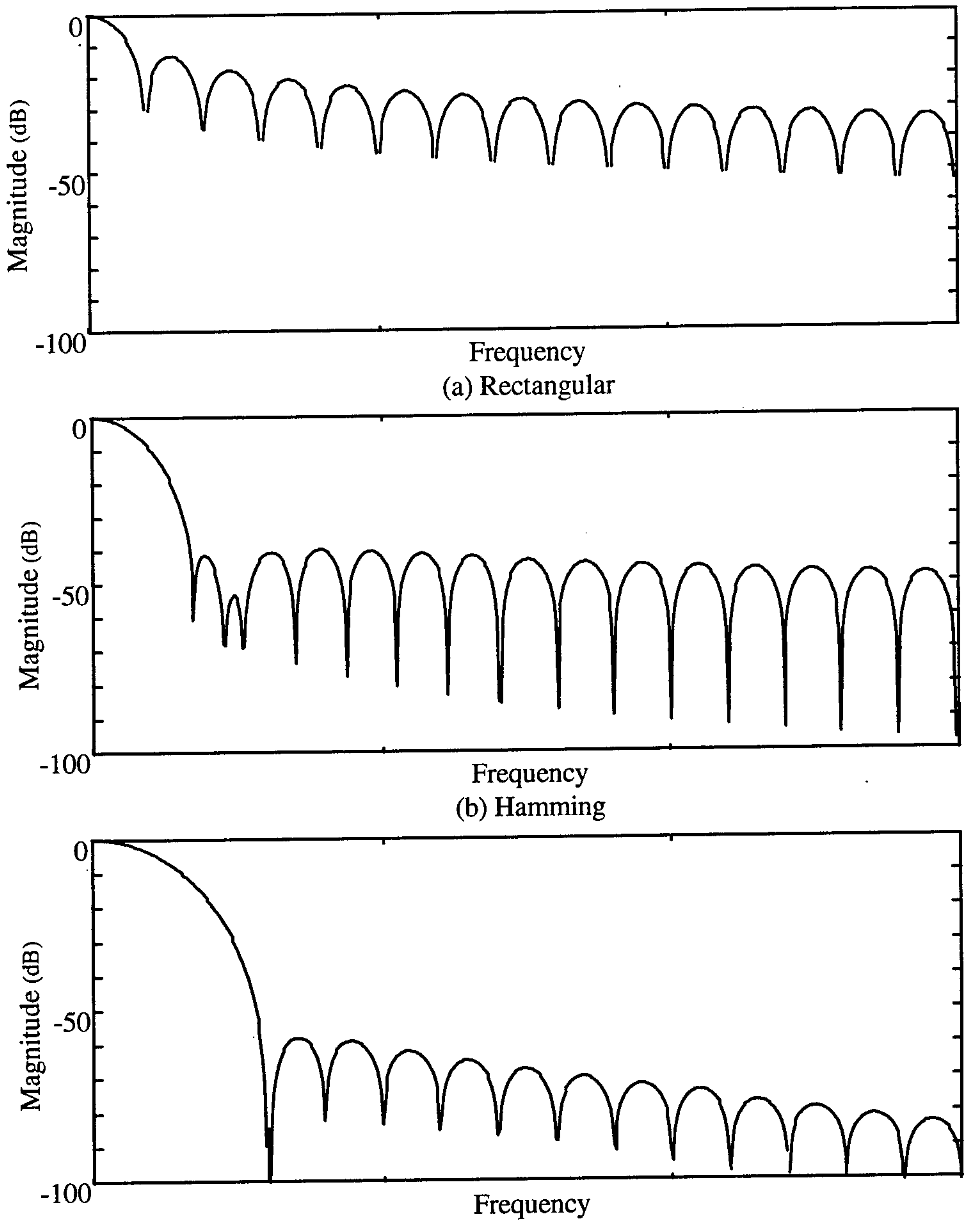


Figure 2.2: Common window functions.



(a) Rectangular

(b) Hamming

(c) Blackman

Figure 2.3: Comparison of the magnitude response for 512 sample length window functions: (a) rectangular, (b) Hamming, (c) Blackman, and

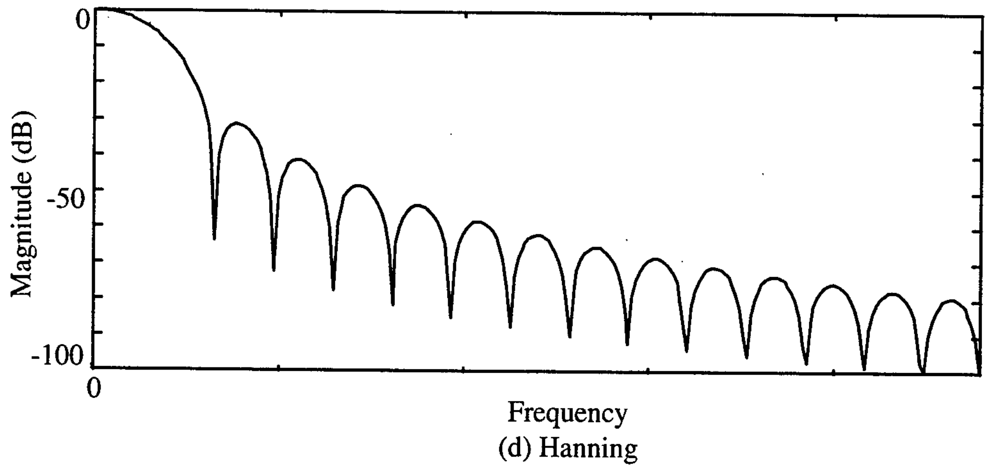


Figure 2.3: Continued, (d) Hanning window function

2.3.2 Filter Bank Analysis

Filter bank analysis is one of the more popular techniques for spectral analysis, even more so with the introduction of wavelets to signal processing. Filter banks are a set of bandpass filters, each capable of analyzing a different range of frequencies of the input signal. It is more flexible than other analysis techniques in that the bandwidths can be varied according to the desired characteristics one needs to find, rather than being fixed for either wideband or narrowband analysis.

2.3.3 Discrete Fourier Transform Spectrum and Analysis

The discrete Fourier transform (DFT) is widely used for finite duration sequences. The DFT is a sequence, as opposed to a function of a continuous variable, which corresponds to samples of the Fourier transform of a signal, equally spaced in frequency.

The direct DFT and inverse DFT of a finite-length sequence $s[n]$ of N samples are, respectively:

$$S[k] = \sum_{n=0}^{N-1} s[n] W_N^{nk} \tag{2.18}$$

and

$$s[n] = \frac{1}{N} \sum_{k=0}^{N-1} S[k] W_N^{-nk} \tag{2.19}$$

where

$$W_N = \exp(-\frac{j2\pi}{N}) = \cos(\frac{2\pi}{N}) - j\sin(\frac{2\pi}{N}) \tag{2.20}$$

The sequence $s[n]$, in (2.19) is represented as a sum of sinusoids of frequencies 0,1,2,...,N-1. Hence the DFT can be interpreted as a frequency analysis of the input signal.

The number of samples chosen for the length of the DFT is inversely proportional to the frequency spacing. As the value of N is increased, we obtain higher frequency resolution but poorer time resolution because the signal properties, averaged over a longer time frame, may not be stationary. As the value of N is decreased, the frequency resolution will degrade but the time resolution will improve since the signal properties will now be averaged over a shorter interval of time.

The power spectral density (PSD) is defined as the discrete Fourier transform of the autocorrelation sequence or it can be represented as the square of the DFT of $s[n]$

$$P(k) = S(k)S^*(k) \tag{2.21}$$

where $P(k)$ is the PSD, $S^*(k)$ is the DFT of $s[n]$ and $S^*(k)$is the complex conjugate of $S(k)$.

The short time spectrum of a signal can be thought of as the product of the spectral envelope, which changes slowly as a function of frequency and the spectral fine structure, which changes rapidly. A voiced speech signal will produce periodic patterns in the spectral fine structure. This is not the case for unvoiced speech. This periodic pattern corresponds to the periodicity of the sound source, produced by the excitation signal. The spectral envelope reflects the resonance and anti-resonance characteristics of the vocal tract.

2.3.4 Short Time Autocorrelation

The short-time autocorrelation function of a signal $s[n]$ is defined as

$$R[m] = \sum_{n=0}^{N-1-|m|} s[n]s[n+m], \qquad m=0,1,2,\ldots,N\text{-}1 \tag{2.22}$$

where N is the frame length. The autocorrelation function preserves the information about a signals harmonic and formant amplitudes as well as its periodicity, while ignoring phase information. The autocorrelation function is used for voiced/unvoiced state estimation, pitch detection, and linear prediction analysis.

2.3.5 Short Time Energy and Magnitude Measures

The short time energy is defined as

$$E = \sum_{m=-\infty}^{\infty} s^2[m]w[m] = \sum_{m=-N/2}^{N/2} s^2[m] \tag{2.23}$$

The short time average magnitude is defined as

$$Mag = \sum_{m=-\infty}^{\infty} |s[m]|w[m] = \sum_{m=-N/2}^{N/2} |s[m]| \tag{2.24}$$

where N is the frame length. The energy measure emphasizes the high amplitudes while the magnitude measure avoids such emphasis. Both methods can be used to clip the speech signal for pitch estimation and separation by removing any low amplitude fluctuations.

2.3.6 Cepstral Analysis

The cepstrum is defined as the inverse Fourier transform of the short-time logarithmic power spectrum $|S(k)|$ of a speech signal $s[n]$:

$$c[n] = DFT^{-1} log_{10}(|S(k)|) \tag{2.25}$$

The cepstrum is a powerful tool for spectral flattening. We can consider the spectrum of a speech signal as the product of the discrete Fourier transform of the glottal excitation $G(k)$ and the transfer function of the vocal tract $H(k)$. We have

$$S(k) = G(k)H(k) \tag{2.26}$$

Then the cepstrum can be written as

$$c[n] = DFT^{-1} log_{10}(|G(k)|) + DFT^{-1} log_{10}(|H(k)|) \qquad (2.27)$$

The cepstrum consist of two components; a slowly varying component which corresponds to the spectrum envelope or the vocal tract filter response and a rapidly varying one which corresponds to the spectral fine structure or the pitch-harmonic peaks. Typically we look at the cepstrum of the DFT of a speech signal and the cepstrum of the LPC coefficients. The DFT cepstral coefficients are calculated from the DFT coefficients using

$$c_n = \frac{1}{N}\sum_{k=0}^{N} log|S(k)| exp(j2\pi kn/N) \qquad (2.28)$$

The LPC cepstral coefficients are calculated from the LPC coefficients using

$$\begin{aligned} \hat{c}_1 &= -a_1 \\ \hat{c}_n &= \begin{cases} -a_n - \sum_{m=1}^{n-1}(1-\frac{m}{n})a_m\hat{c}_{n-m} & 1 < n \le p \\ -\sum_{m=1}^{p}(1-\frac{m}{n})a_m\hat{c}_{n-m} & p < n \end{cases} \end{aligned} \qquad (2.29)$$

where the a_i represent the LPC coefficients.

2.3.7 Zero-crossing Measure

The short time average zero-crossing rate is an efficient technique to aid in the determination of the voicing state of a segment of speech, to detect the periodicity of the sound source or to estimate the fundamental period. A zero crossing occurs whenever the signal waveform crosses the time axis. The zero crossing rate is the number of times a zero crossing occurs within a finite time duration. It can also be thought of as the number of times a sequence of discrete data changes sign. Formally defined, the short time zero crossing rate measure of an *N*-length interval of data is given by

$$Z_s[m] = \sum_{n=m-N/2}^{m+N/2} \frac{\left|sgn\{s[n]\} - sgn\{s[n-1]\}\right|}{2} w[m-n] \tag{2.30}$$

where

$$sgn\{s[n]\} = \begin{cases} +1, & s[n] \geq 0 \\ -1 & s[n] < 0 \end{cases} \tag{2.31}$$

and *w*[*m-n*] is the windowing function.

2.3.8 Mel-Cepstrum

The mel-cepstrum coefficients are cepstral coefficients calculated using a mel-frequency scaling. A block diagram of the method used to obtain the coefficients is shown in Figure 2.4. The input signal is windowed using a Hanning-weighted window function. A shallow high-pass filter is used for pre-emphasis. The log-magnitude spectrum of the signal is then compressed using a triangle weighting function. The center frequencies of the triangle weighting functions are spread across the spectrum such that both the center frequencies and the bandwidth increase with frequency. This spacing models the frequency sensitivity of a human auditory system, consistent with the mel-scale. The coefficients can be calculated using [54]

$$\tilde{c}_n = \sum_{k=1}^{K} \tilde{S}_k \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad n = 1,2,\ldots,L, \tag{2.32}$$

where $\tilde{S}_k$ are the power coefficients calculated from the log-magnitude spectrum of the signal, K is the number of frequency bins and L is the desired length of the cepstrum.

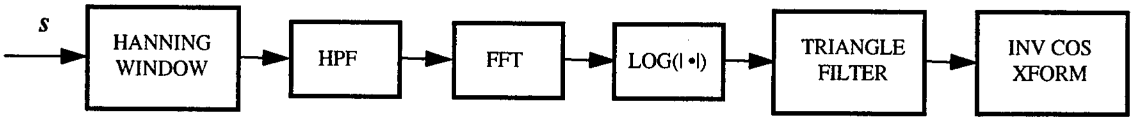


Figure 2.4: Block diagram on the calculation of the mel-cepstrum coefficients.

2.4 Constrained Optimization

Optimization is the process of finding the best solution to a given problem. Typically this involves finding the maximum or minimum of an objective function of n variables, $f(x_1,\dots,x_n)$ where n is an integer greater than zero. When some or all the variables of the objective function have restrictions or bounds, the optimization is said to be constrained. For general linear functions, subjected to linear equality constraints, optimization problems reduce to the form

$$\min \sum_{j=1}^{n} c_j x_j \qquad (2.33)$$

subject to the constraints

$$\sum_{j=1}^{n} a_{ij} x_j \leq b_i \qquad (i = 1,\dots,m) \qquad (2.34)$$

where c_j and a_{ij} are constants or known data and the x_j are the variables. The study of such problems of this form are known as linear programming.

When we are dealing with nonlinear objective functions, optimal solutions become more difficult with local minima posing as erroneous solutions. For the general nonlinear programming problem of the form

$$\begin{array}{lll} \text{minimize} & F(x) & \\ \text{subject to} & g_i(\mathbf{x}) \le b_i & (i = 1,\ldots,m) \\ & x_j \ge 0 & (j = 1,..n) \end{array} \tag{2.35}$$

Second derivatives for general nonlinear objective functions are relatively difficult to obtain. At times it may be required that the first derivative (the gradient) must be perturbed in order to obtain reasonably accurate approximations of the second derivatives. Several methods, including the Gauss-Newton method and the Levenberg-Marquardt (LM) technique can be effective in determining the proper steps toward an optimal solution.

2.5 Pattern Recognition

Pattern recognition is a decision-making process. It is the formulation of a decision based on the classification of a set of measured features. Typical applications can be lumped into two categories, either classification of waveforms or classification of geometric figures. Here, we are dealing with the classification of a waveform, the waveform being a segment of speech. Decisions are made with the use of a continuous pattern recognition system that extract a set of features or discriminants from a segment of speech. These discriminants will then be presented to a classifier that will then make a decision based on the type of speech present.

Figure 2.5 shows an example of a generic pattern recognition system. Here, an *n*-dimensional sampled waveform is transformed to an *m*-dimensional feature waveform, where $m << n$. This feature vector is then passed through a classifier that makes a decision on the class in which the feature vector most closely matches. The training data set of feature vectors are used to train a classifier to distinguish between different classes. The training feature vectors are presented to the classifier along with the *a priori* knowledge that they belong to a particular class. This is how the classifier *learns* to make correct decisions.

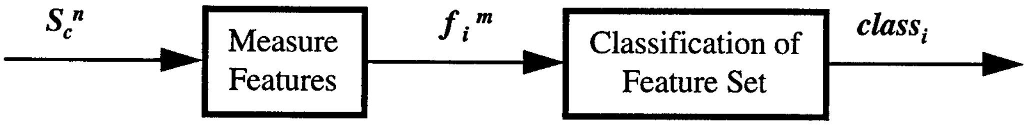


Figure 2.5: Generic pattern recognition system.

The difficulty of a pattern recognition system arises in choosing the optimal features to describe a waveform and choosing the best classifier to discriminate between classes based on the feature set. The optimal classifier is chosen based on knowledge of the class statistics of the feature vectors. We must study the underlying probability density functions of the feature vectors to find the proper discriminant function. Classical approaches to designing a pattern recognition system is the problem of estimating density functions in a multi-dimensional space and then dividing the space into regions of classes.

The Bayes classifier is the best classifier to minimize the probability of classification error when the distributions of the feature vectors are known. However, depending on the density functions and the dimensionality of our feature vector, implementation of the Bayes classifier can become complex and may not be the most practical. Therefore, designers are often led to consider simpler, parametric and nonparametric classifiers.

Parametric classifiers are based on assumed mathematical forms of either the density functions or the discriminant functions. Typically, the parameters that describe the density function, such as the mean and variance, are estimated from available samples. Given a finite number of samples, these parameters and consequently these classifiers, become random variables. The classification error also becomes a random variable and is biased with a variance. Therefore, the number of samples becomes important in the performance of the classifier and its design. An example of a parametric classifier is the quadratic classifier.

When it is not possible to assume any parametric structure for the density functions, we must use nonparametric techniques. In nonparametric approaches, the density function is estimated from a given number of neighboring samples. These estimates, which are used in classification, are less reliable, with a larger bias and variance than the parametric estimates. Two examples of nonparametric techniques for classification are the *k-nearest neighbor* and Parzen window classifiers. An in-depth discussion of these approaches is provided in the next chapter

3. CO-CHANNEL SPEAKER SEPARATION

In this chapter we develop a method to simultaneously separate two overlapping speech signals recorded onto a single channel. This process involves three different filtering techniques. Two of these techniques are traditional while the third is innovative and unique. We also present and develop a technique to predict the voicing state of each speaker present in the co-channel signal, a method to measure the pitch contour of two overlapping speech signals, and an algorithm to perform harmonic selection.

The possible voicing state combinations of co-channel speech can be labeled as (desired speaker / interfering speaker): *silence*, *voiced/voiced*, *voiced/unvoiced*, *unvoiced/voiced*, and *unvoiced/unvoiced*. We have developed a novel approach to predict the voicing state of each speaker present in a co-channel signal. This technique is based on a pattern recognition approach. Features measured from the segments of co-channel speech are classified using a parametric classifier. A nonparametric classifier is also presented.

For overlapping voiced segments of speech, we have developed a constrained nonlinear least squared optimization algorithm that simultaneously estimates the spectral characteristics of two voiced speech signals. The spectral characteristics include the amplitude, frequency and phase of the spectral harmonics of each speaker, based on a sinusoidal representation for voiced speech. Prior to optimization, the fundamental frequencies of these two overlapping segments are measured and then used to perform harmonic selection. Harmonic selection is the process of identifying a speaker's spectral harmonics in a co-channel speech spectrum.

The first section of this chapter presents our preliminary research into the study of the intelligibility of a speech signal and how it is affected by enhancement or suppression using an excitation-modulation speech model. This provides insight into the limitations of suppression and enhancement systems for co-channel speech and provides the motivation that directed us to a system that simultaneously estimates both signals. The remaining sections present the theory on constrained nonlinear least squared optimization applied to the separation of overlapping speech, development of a voicing state determination algorithm, joint pitch estimation algorithm, a harmonic selection algorithm and speech reconstruction.

3.1 Preliminary Research

Our investigation of preliminary speaker separation systems provided the motivation to develop a system that simultaneously estimates overlapping speech signals. Before presenting this material, it is important to first investigate how different processing techniques affect speech intelligibility. This provides insight into which parts of a speech signal are crucial to separation and intelligibility. Also it demonstrates the relative importance between the excitation signal and the spectral envelope and the need in a speaker separation to not only estimate the amplitude of the spectral harmonics of signal, but also the phase and center frequency.

3.1.1 Harmonic Enhancement

Given an excitation-modulation speech model, similar to the one presented in Section 2.1, a speech signal can be deconvolved into two parts, an excitation or driving function and a vocal tract filter. The excitation function contains the spectral fine structure or harmonic characteristics of a speech signal. The vocal tract filter models the spectral envelope or modulation part of a speech signal. The excitation signal can be represented as white noise or as a train of impulses produced at a given repetition rate. This rate is the fundamental or pitch frequency.

Hanson and Wong [10] outlined several scenarios using this model for speech. We have performed tests on these scenarios to gain greater insight into co-channel speech. These tests, outlined in Figure 3.1, consist of simultaneously separating out the excitation signal from the spectral envelope of the co-channel signal (*s*+*i*) and the excitation signal from the spectral envelope of the signal (*s*) alone. The signal (*s*) can be thought of as the desired speech signal and the interference (*i*) as the undesired or interfering speech signal. The spectral envelope for both inputs is modeled as an autoregressive (AR) process or all-pole filter using a forward linear prediction estimate. Using forward linear prediction coding, we can represent the estimate of an AR model as

$$\hat{s}^f[n] = -\sum_{i=1}^{p} a_i^f s[n-i] \tag{3.1}$$

The residual error signal then becomes

$$e^f[n] = s[n] - \hat{s}^f[n] \tag{3.2}$$

Equation (3.2) represents the excitation signal to the all-pole LPC filter to produce speech. We can write a similar equation for the co-channel signal (s_c) as

$$\hat{s}_c^f[n] = -\sum_{i=1}^{p} a_{ci}^f s_c[n-i] \tag{3.3}$$

where

$$s_c[n] = s[n] + i[n] \tag{3.4}$$

The residual error signal is given by

$$e_c^f[n] = s_c[n] - \hat{s}_c^f[n] \tag{3.5}$$

Two outputs are produced at a signal to interference ratio (SIR) of -6 dB. The first output, s_1, is obtained by driving the LPC synthesis filter derived from $s+i$ with the excitation signal of s:

$$s_1 = w_{s+i}^f[n] \otimes e_s^f[n] \tag{3.6}$$

where $w_{s+i}^f[n]$ is the impulse response of the co-channel vocal tract filter model. The second output, s_2 is obtained by driving the LPC synthesis filter derived from s with the excitation signal of $s+i$.:

$$s_2 = w_s^f[n] \otimes e_{s+i}^f[n] \tag{3.7}$$

and similarly $w_s^f[n]$ is the impulse response of the signal's vocal tract filter model.

Several speech signals were tested. An example of output signals s_1 and s_2 are shown in Figure 3.2 and Figure 3.3 respectively. This example was produced from a mixture of male and female speech signals with different average pitch frequencies. The male speech was labeled as the signal and the female speech was selected as the interference. Both output signals are compared to the original co-channel signal for improvement in intelligibility.

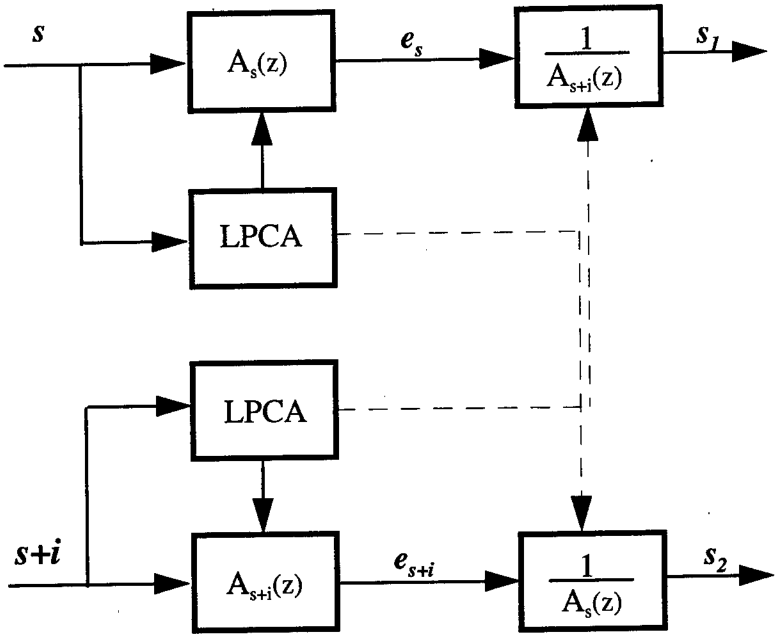
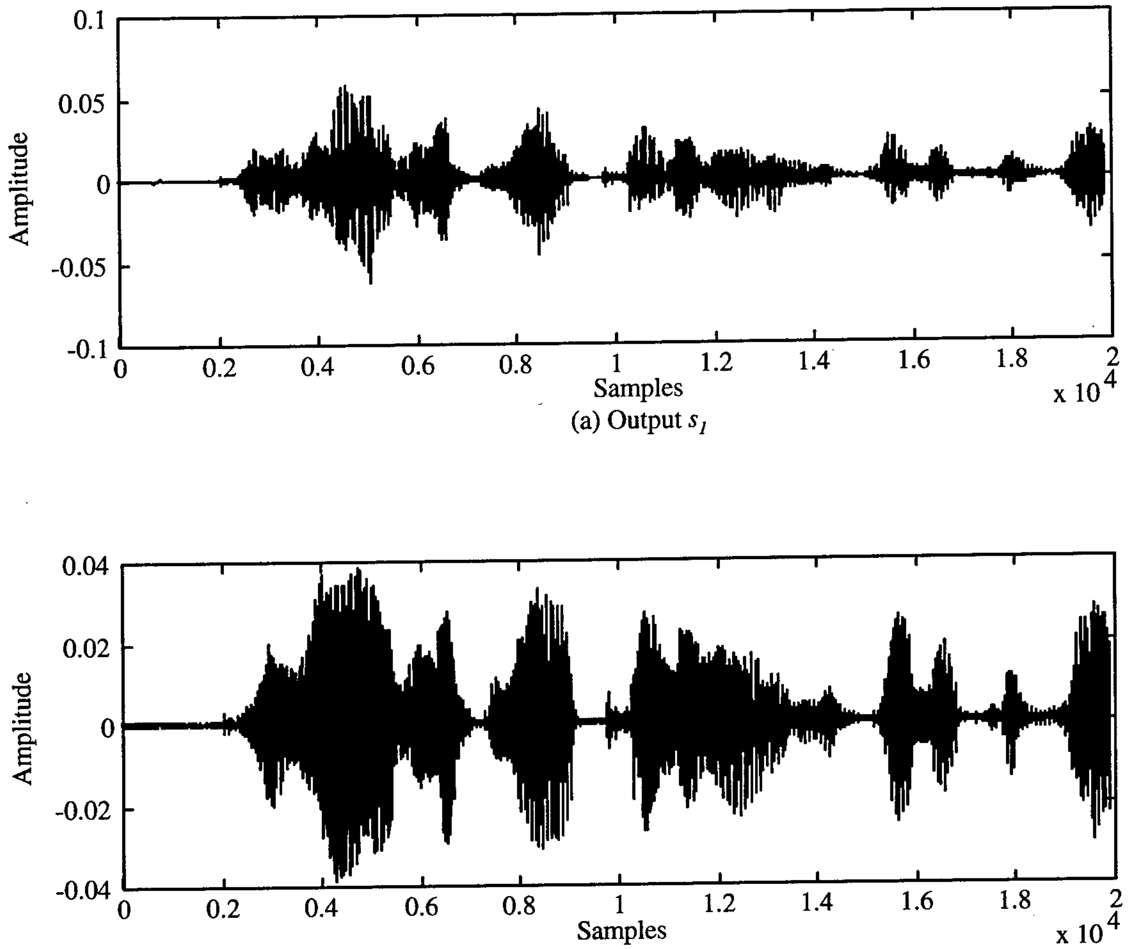


Figure 3.1: Block diagram showing the excitation of the signal's LPC synthesis filter with the co-channel excitations signal and excitation of the co-channel LPC synthesis filter with the signal's excitation signal.

Results of tests using speech signals taken from the TIMIT database are given below. It is well known that exciting the envelope of a speech signal with even random noise will produce “whispered” but intelligible speech (s_1). However, it was found that when the desired excitation signal is used to excite the LPC synthesis filter derived from the co-channel speech (s_2), it provided a slight improvement in intelligibility over that of s_1.

These results lead us to believe that improvement in intelligibility is possible when the spectral fine structure of a speech signal is accurately modeled and then used to enhance a corrupted estimate of the spectral envelop of the desired signal. Exciting the LPC synthesis filter of the co-channel speech with the driving function from the desired speech signal will enhance the spectral characteristics (phase and amplitude) of the desired speaker that are contained in the co-channel spectral envelope. However, we must also consider the effect this will have on the spectral characteristics of the interfering speaker's spectral components.

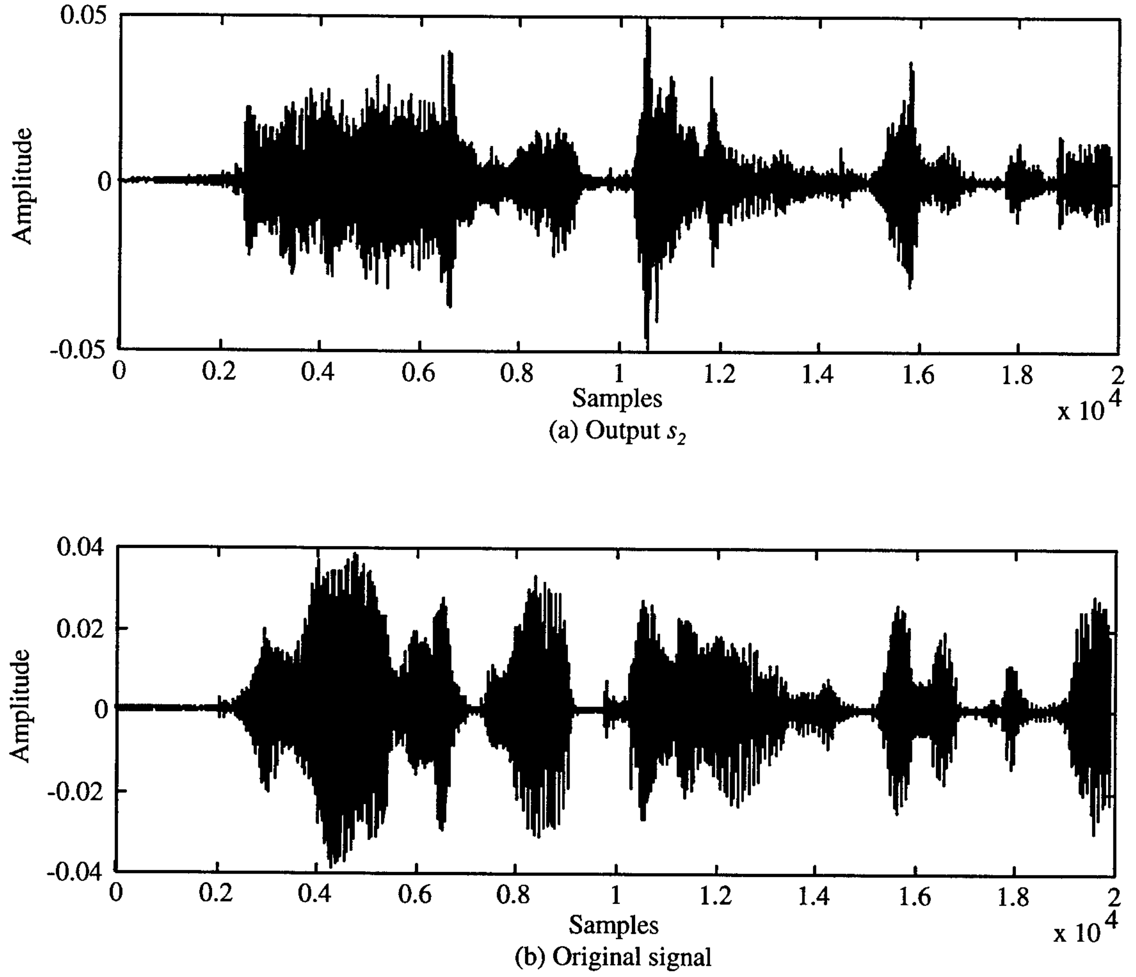
In two separate tests both Perlmutter et al. [43] and Hanson and Wong [10] found that the intelligibility of co-channel speech using enhancement techniques had intelligibility scores which were consistently lower for the processed (enhanced) speech when compared to the unprocessed co-channel speech. This can be attributed to the fact that traditional methods that were used to enhance the desired speaker's spectral characteristics in the corrupted speech also enhanced adjacent harmonic components of the interfering speech signal. Since, in their test the signal power of the interfering speech signal was stronger than the desired speech signal, this method tends to produce even greater interference than if the co-channel signal was left unprocessed. Using simultaneous estimation of both signals, we will show that processing improves the intelligibility of both signals at varying SIRs.



(a) Output s_1

(b) Original signal

Figure 3.2: Example of the effects of speech enhancement. (a) Plot of output s_1 (excitation of co-channel LPCs with signal's driving function) and (b) plot of original speech signal.



(a) Output s_2

(b) Original signal

Figure 3.3: Example of the effects of speech enhancement: (a) plot of output s_2 (excitation of original LPCs with the co-channel driving function) and (b) plot of original speech signal.

3.1.2 Harmonic Suppression

Harmonic suppression of co-channel speech is a process by which the position of the spectral harmonics of the interfering speech signal are estimated and then suppressed from the co-channel spectrum while the harmonics of the desired speech signal are left unprocessed. A well known method of performing harmonic suppression is the method of spectral magnitude subtraction. Here an estimate of the magnitude spectrum of the interfering signal is subtracted from the magnitude spectrum of the co-channel signal leaving an estimate of the magnitude spectrum of the desired signal.

In cases in which speech has been corrupted by another speech signal, it has been shown that at low SIR, magnitude suppression is the preferred method [10]. Harmonic suppression works well for interference in which the spectrum of the interfering signal can be accurately estimated. The difficulty lies in estimating the magnitude spectrum of the interfering signal. For voiced speech, this requires precise knowledge of the shape, amplitude, and position of the interfering spectral harmonics. To test this premise we have assumed *a priori* knowledge of these interference characteristics to accurately test the performance of harmonic suppression for speech intelligibility. This provides an upper limit on performance.

In the previous section, we investigated the merit of speech enhancement on speech intelligibility by exciting an LPC synthesis filter derived from the co-channel speech with the excitation signal of the desired speech. It was found that while this technique enhanced the spectral characteristics of the desired speech, in many cases it also enhanced the spectral characteristics of the interfering speech, providing no improvement in intelligibility over the unprocessed co-channel speech. Here, we investigate suppressing the harmonics of the interfering speech, while still trying to preserve the spectrum of the desired speech using harmonic magnitude suppression.

First, we must consider the limitations inherent to magnitude spectral suppression. Let us define $S(k)$ as the discrete Fourier transform (DFT) of the desired speech signal $s[n]$, $I(k)$ as the DFT of the interfering speech signal $i[n]$, and $S_c(k)$ as the DFT of the co-channel speech $s_c[n]$. From the definition of the Fourier transform, it can be clearly seen that when

$$s_c[n] = s[n] + i[n] \tag{3.8}$$

then

$$S_c(k) = S(k) + I(k). \tag{3.9}$$

These terms represent phasors, with a given magnitude and phase. An inherent assumption is that when signals are combined in the time domain, then the sum of the spectra are related in the magnitude spectrum domain. However, this is not the case with

$$|S(k) + I(k)| \neq |S(k)| + |I(k)|. \tag{3.10}$$

That is, the spectrum of the sum of two signals in the time domain is *not* equivalent to the sum of the spectra of each signal in the magnitude spectral domain. As can be seen in Figure 3.4, the magnitude spectrum of the sum of two signals may, and usually will be significantly different than the sum of the magnitude spectrum of the signals separately. The solid line plot in this figure shows the magnitude spectra of the co-channel speech signal when the signals are combined in the time domain. The dashed line plot is a result of when the magnitude spectra are combined in the frequency domain.

Another drawback to magnitude suppression is that the phase of the signal is required for resynthesis. Most researchers have substituted the phase of the co-channel speech for the phase of the desired speech. However, the phase of the co-channel signal will more closely represent the phase of the stronger or interfering signal, not the desired, weaker signal. Given an understanding of these limitations, we can now test the effectiveness of magnitude suppression on reducing co-channel speech interference. We start with an ideal situation and work towards a more realistic scenario.

The first test, an ideal case, is depicted in Figure 3.5 where we assume the magnitude spectrum of the interference is known completely. The interference and the co-channel speech signals are treated separately on a frame by frame basis. First, we calculate the magnitude spectrum of the interference and co-channel speech separately for each frame including voiced and unvoiced interference. Next, we subtract the magnitude spectrum of the interference from the magnitude spectrum of the co-channel signal. We then reconstruct the desired speech signal by taking the inverse DFT (IDFT) of this difference along with the phase of the co-channel speech. An example of the recovered speech signal, $\hat{s}$ shown in Figure 3.6 is compared to the original speech signal. The plot in Figure 3.6(a) is the recovered speech signal and the plot in Figure 3.6(b) is the original speech signal.

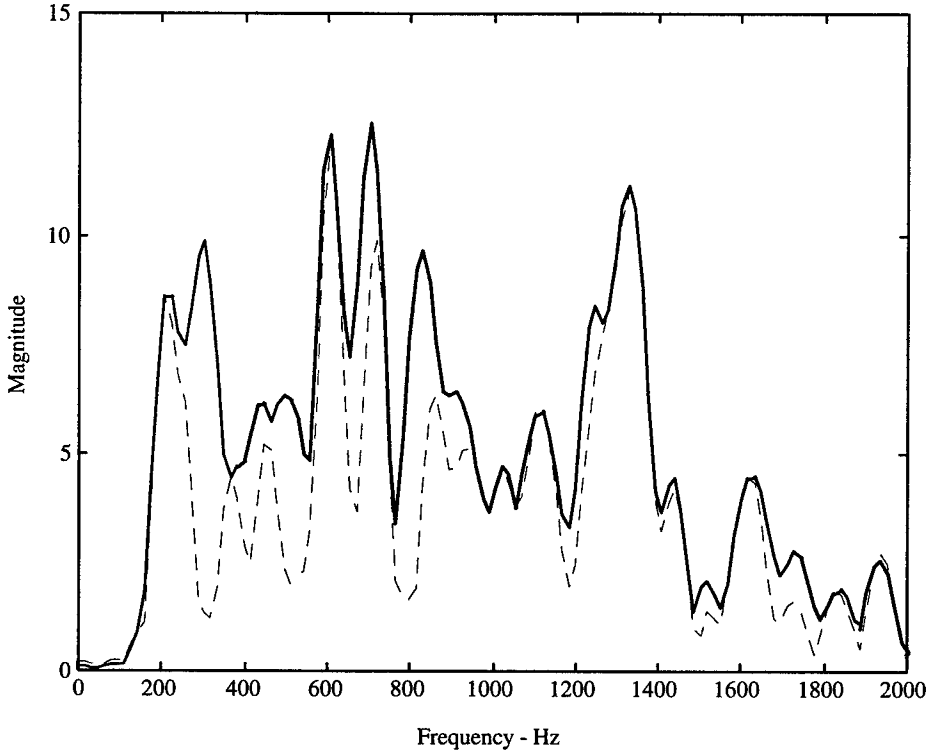


Figure 3.4: Solid Line Plot - spectral magnitude of co-channel speech combined in the time domain. Dashed Line Plot - sum of individual magnitude spectra.

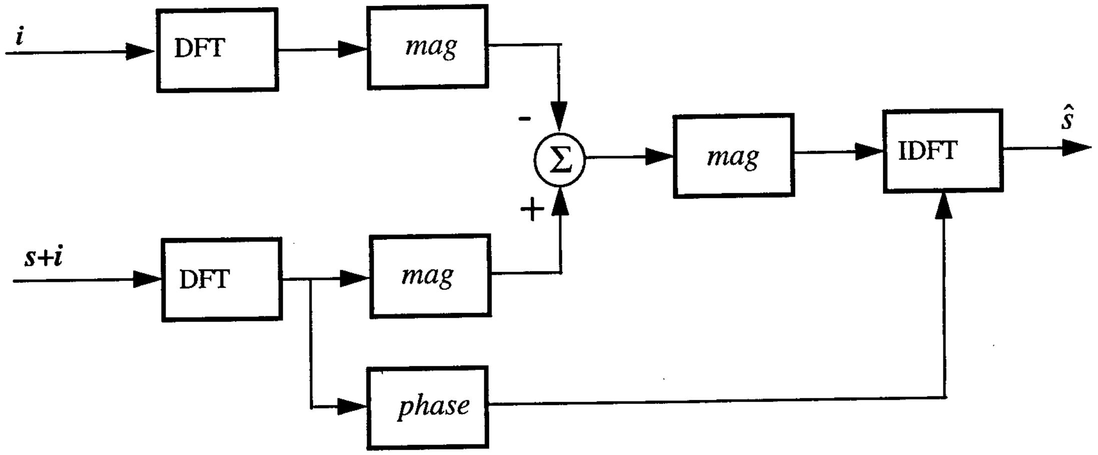
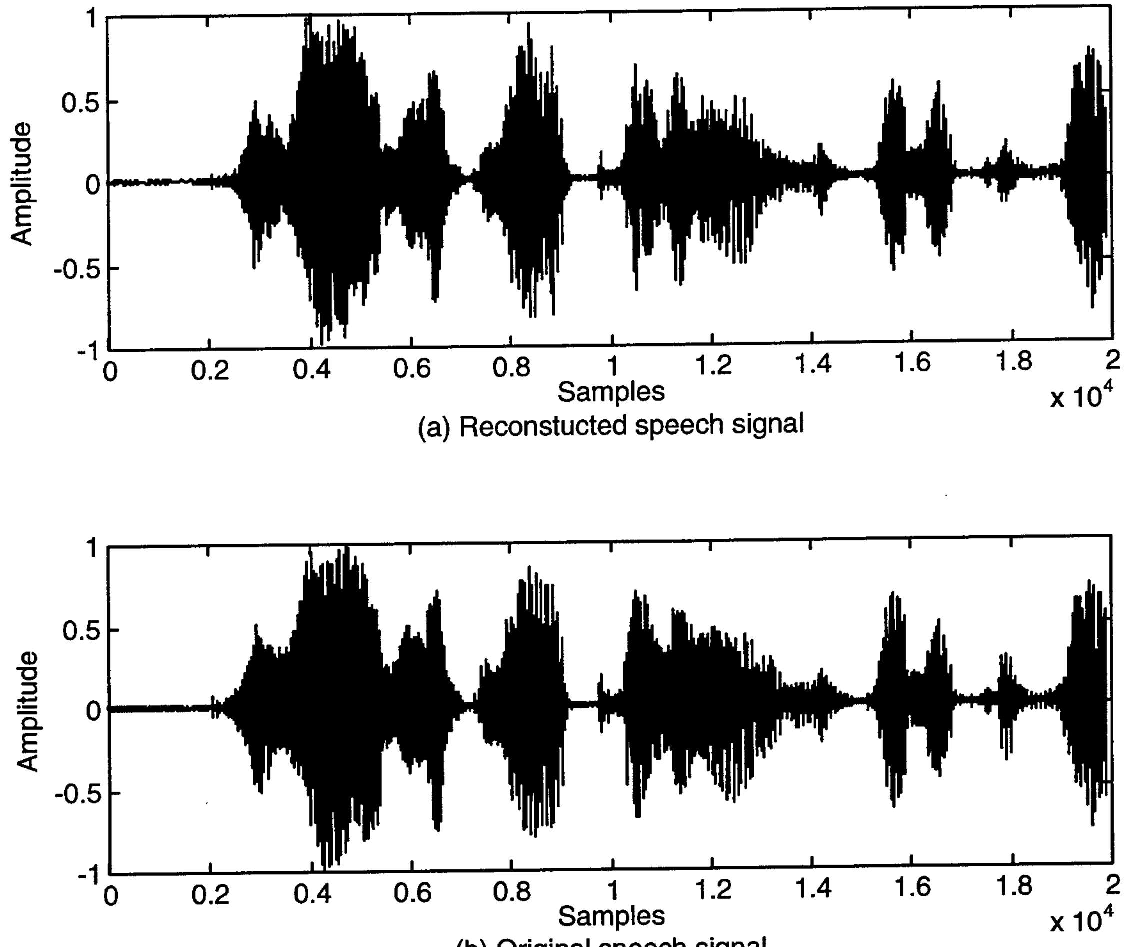


Figure 3.5: Spectral Magnitude Subtraction - ideal case.

When the interference magnitude spectrum is known, spectral subtraction provides a very accurate estimate of the original signal. Informal listening clearly demonstrates that this type of harmonic suppression is an improvement over the harmonic enhancement method previously presented. However, this is an ideal situation, providing only the best case scenario.

A slightly more realistic approach is depicted in Figure 3.7 with results in Figure 3.8. Here we perform harmonic suppression using magnitude spectral subtraction only when the interfering speech signal is voiced and we employ a lowpass filter (LPF) with a cut-off frequency of 4 kHz when the interfering speech is unvoiced. Informal listening tests have shown that intelligibility of this method is very close to the level obtained by the previous method. This leads us to a significant conclusion that it is not necessary to estimate the noise spectrum when the interference is unvoiced. A lowpass filter performs sufficient suppression of unvoiced interference on a voiced speech signal. However, we do need to have some way of knowing the voicing state of both the desired signal and the interference signal.

An even more realistic approach to harmonic suppression, one which uses less *a priori* information is to estimate the pitch of the interfering signal and use this to estimate the locations of the harmonics of the voiced interference. Given these locations a comb filter, with uniform amplitude is implemented to suppress the harmonics of the interference. A more accurate spectrum would be to estimate the amplitude of each harmonic by sampling the co-channel spectra at the harmonic's center frequency [10]. Clearly this can only be as good, but no better than the ideal spectral suppression method described above. The limited success of these methods is mainly due to the difficulty in estimating the center frequency, amplitude and phase of each harmonic.



(a) Reconstucted speech signal

(b) Original speech signal

Figure 3.6: Results of Spectral Magnitude Subtraction when the magnitude spectrum of the interference is known and the phase of the co-channel speech is used: (a) the reconstructed speech signal, (b) the original speech signal.

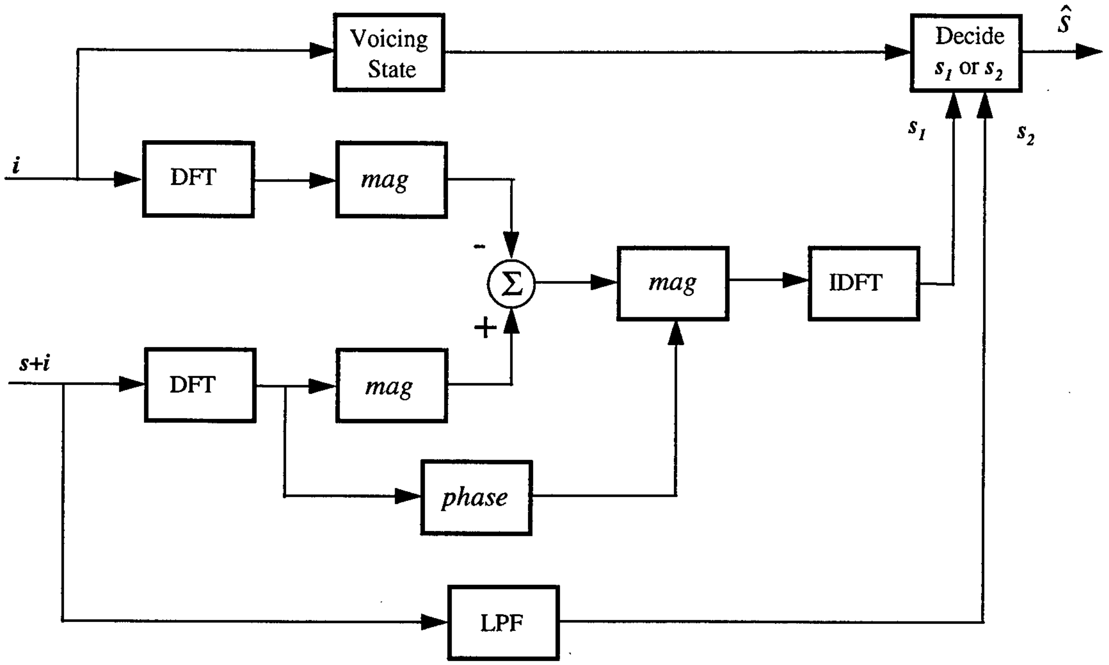
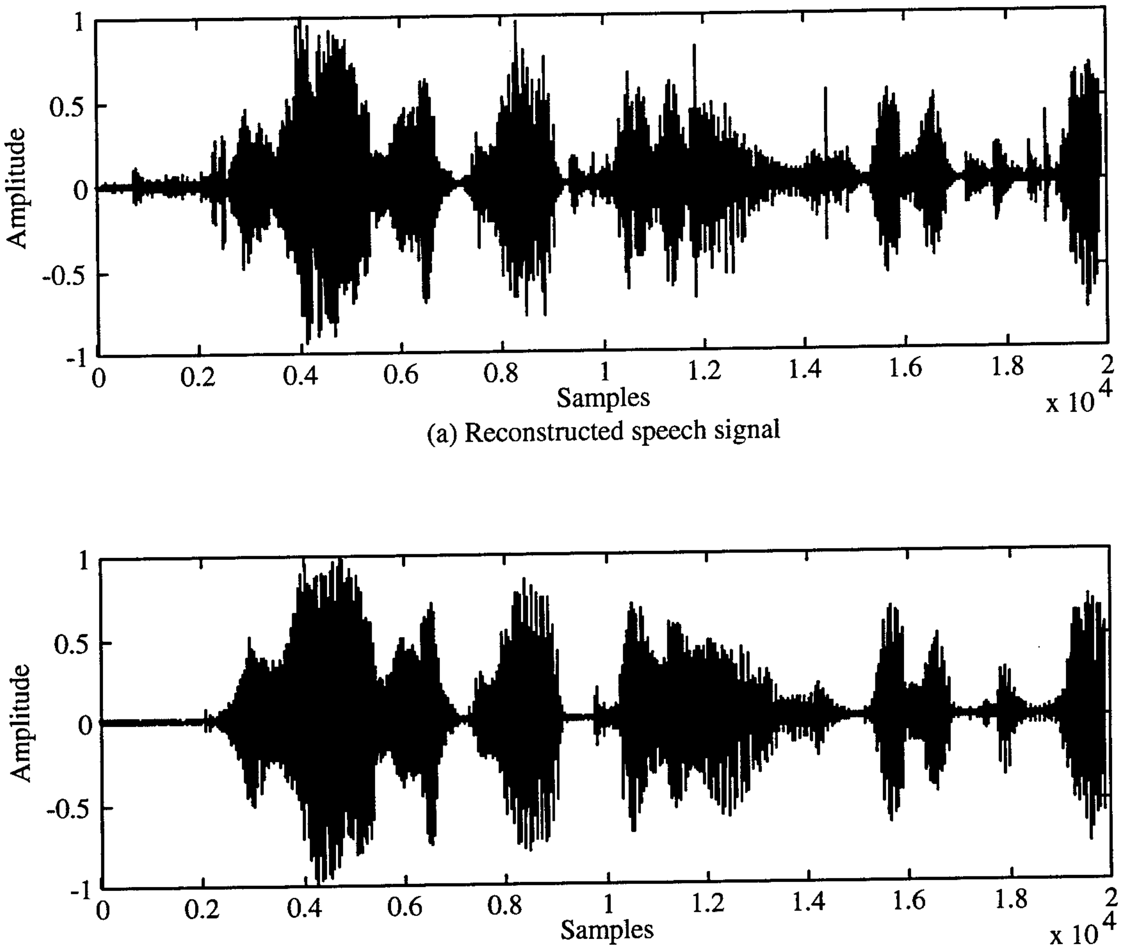


Figure 3.7: Spectral Magnitude Subtraction with LPF - spectral magnitude subtraction is used for voiced interference and an LPF is applied when interference is unvoiced.



(a) Reconstructed speech signal

(b) Original speech signal

Figure 3.8: Results from Spectral Magnitude Subtraction using an LPF for unvoiced interference: (a) reconstructed speech signal, and (b) original speech signal.

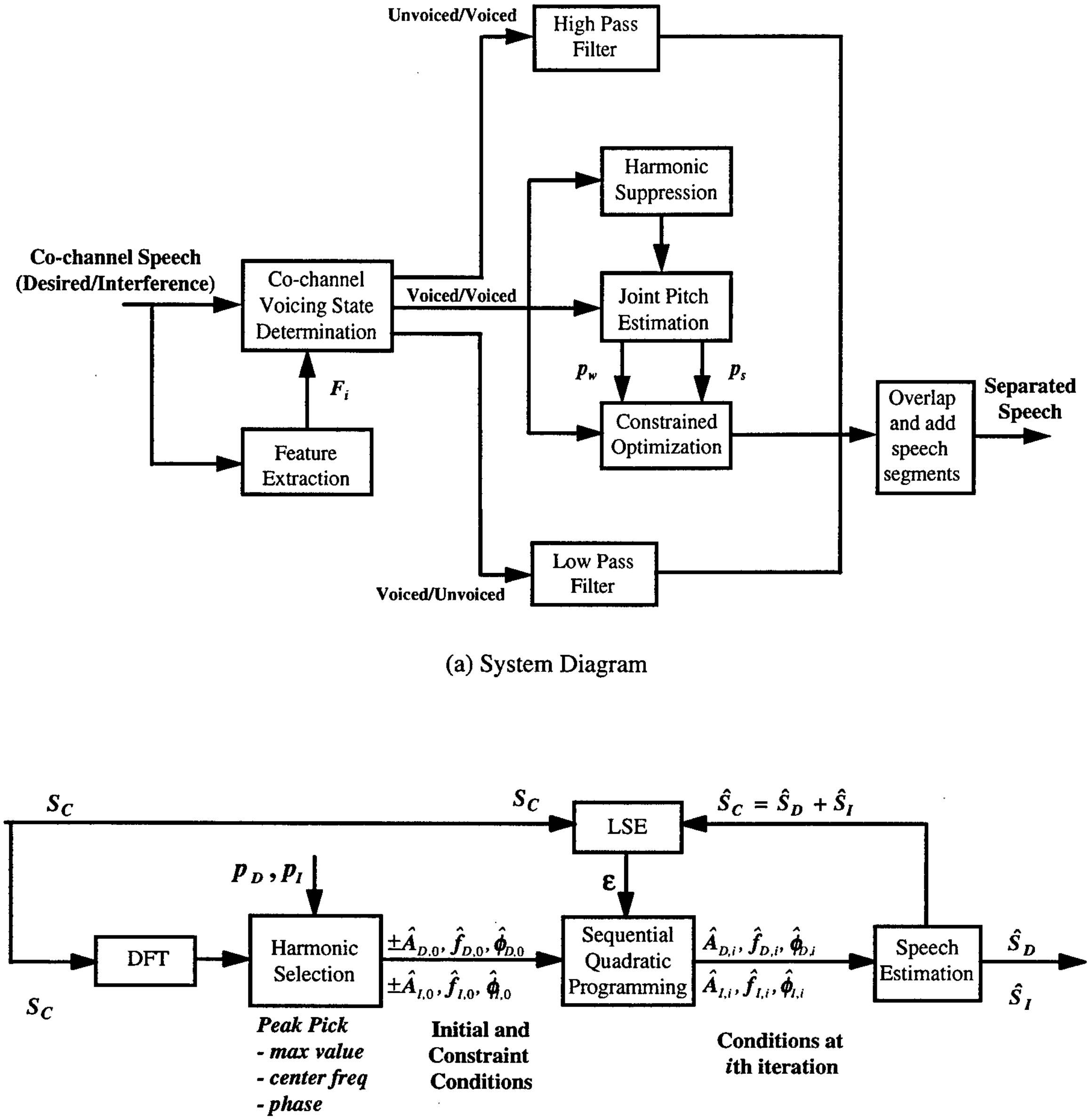
3.1.3 Current Research

Our current research has advanced the work of our preliminary research, utilizing not one, but several processing methods, based on the voicing state of each speaker, to separate overlapping speech signals. Our method requires accurate estimates of the voicing state of the speakers present and accurate measurements of the pitch frequencies. With this information, our technique simultaneously separates the spectral characteristics of each speaker.

Our system, outlined in Figure 3.9, provides the framework to successfully separate overlapping speech signals. Referring to Figure 3.9, the co-channel speech signal, $s_C = s_D + s_I$, which is the sum of the desired and interfering speech signals, is broken into segments or frames 30 msec. in duration. Each frame is processed separately. A pre-processor extracts speech characteristics, $\overline{F_i}$. This information is used to predict the voicing state of co-channel speech. Possible states for each speaker are voiced, unvoiced or silence. The predicted voicing state for each speaker is used to decide which separation technique is applied. Co-channel speech that is all unvoiced or silent is left unprocessed.

Referring to the top branch of Figure 3.9, when the desired speech is unvoiced and the interfering speech is voiced, the co-channel speech signal is highpass filtered to remove the effects of the interfering speech signal. The residual signal is retained for reconstruction of the interfering signal. When the desired speech is voiced and the interfering speech is unvoiced, the co-channel speech is lowpass filtered to remove the effects of the interfering speech. Again, the residual signal is retained for reconstruction of the interfering signal. Constrained nonlinear least squared optimization is used when both the desired and interfering speech signals are voiced. An estimate of the pitch frequency of the stronger speaker is used to suppress the harmonics of that speaker. The pitch frequency of the weaker signal is then estimated from this resulting signal. This pitch information is used to estimate the harmonics of both speakers which is then used to initialize the optimization routine.

Following the constrained nonlinear least squared optimization branch, the co-channel speech signal is passed through a discrete Fourier transform. Initial values of the spectral harmonic peaks of both signals are estimated from this signal using estimates of the fundamental frequencies. These values are also used to provide constraints to the optimization variables. These variables include the amplitude, phase and center frequency of each spectral harmonic. A sequential quadratic programming algorithm adjusts these variables to their optimal values. This optimization routine minimizes the squared error between the original co-channel speech segment and the sinusoidal representation of the co-channel segment. Once a minimum has been reached and the desired and interfering speech segments have been found, an overlap and add technique is used to reconstruct the speech segments into intelligible speech signals. Segments of speech in which both speakers are silent or both speakers are unvoiced are left unprocessed.


Unvoiced/Voiced
High Pass Filter
Harmonic Suppression
Co-channel Speech (Desired/Interference)
Co-channel Voicing State Determination
Voiced/Voiced
Joint Pitch Estimation
p_w
p_s
F_i
Constrained Optimization
Overlap and add speech segments
Separated Speech
Feature Extraction
Low Pass Filter
Voiced/Unvoiced
s_C
s_C
LSE
$\hat{s}_C = \hat{s}_D + \hat{s}_I$
p_D, p_I
ε
DFT
Harmonic Selection
$\pm\hat{A}_{D,0}, \hat{f}_{D,0}, \hat{\phi}_{D,0}$
$\pm\hat{A}_{I,0}, \hat{f}_{I,0}, \hat{\phi}_{I,0}$
Sequential Quadratic Programming
$\hat{A}_{D,i}, \hat{f}_{D,i}, \hat{\phi}_{D,i}$
$\hat{A}_{I,i}, \hat{f}_{I,i}, \hat{\phi}_{I,i}$
Speech Estimation
$\hat{s}_D$
$\hat{s}_I$
s_C
Peak Pick
- max value
- center freq
- phase
Initial and Constraint Conditions
Conditions at *i*th iteration

(a) System Diagram

(b) Constrained nonlinear least squared optimization

Figure 3.9: Simultaneous Adaptive Speaker Separation

3.2 Constrained Nonlinear Least Squared Optimization

We have shown in Section 3.1.2 that with an accurate estimate of the magnitude spectrum of the interfering speech, spectral magnitude subtraction improves the intelligibility of speech in the presence of voiced interference. The major drawback to implementation of such a system is the need for an accurate estimate of the magnitude spectrum of the interfering signal. That is, we must have an accurate estimate of the center frequency, the spectral amplitude, phase, and the proper shape of each harmonic of the interfering spectrum. Here, we develop a method to simultaneously estimate the center frequency, amplitude and phase of all significant harmonics for both signals present in an overlapping voiced co-channel speech segment.

The speech production system model can be represented as the output of a vocal tract filter excited by a train of impulses for voiced speech. Given this model, a voiced speech waveform can be represented as a sum of sine waves, each with a time-varying amplitude, frequency and phase. Our speech signal $s[n]$can be written as

$$s[n] = \sum_{k=1}^{M} a_k[n] cos(\theta_k[n]) \tag{3.11}$$

where the amplitudes are denoted by $a_k[n]$ and the phase terms by $\theta_k[n]$. We can simplify the phase function by assuming the speech signal is stationary so that each sine function can accurately be represented by a fixed phase, θ_k and a fixed amplitude a_k in a given time interval [41]. Our representation is then written as

$$s[n] = \sum_{k=1}^{M} a_k \, cos(\omega_k n + \phi_k) \tag{3.12}$$

where

$$\theta_k[n] = \omega_k n + \phi_k \tag{3.13}$$

and ϕ_k is the phase offset measured relative to the beginning of the segment of data (i.e. n=0).

For our application, we sum the two speech segments, in the time domain, to produce the co-channel speech signal

$$s_c[n] = s_a[n] + s_b[n] \tag{3.14}$$

where

$$s_a[n] = \sum_{k=1}^{M_a} a_k \, cos(\omega_{a,k} n + \phi_{a,k}) \tag{3.15}$$

$$s_b[n] = \sum_{k=1}^{M_b} b_k \, cos(\omega_{b,k} n + \phi_{b,k}) \tag{3.16}$$

Again, we have assumed a fixed frequency and amplitude within a given time interval. This is equivalent to the assumption the signal exhibits quasi-stationary spectral properties throughout the length of the speech segment. Physically, this implies that the vocal cords and vocal tract characteristics are fixed during our time interval.

We can substitute (3.15) and (3.16) into (3.14) to give

$$s_c[n] = \sum_{k=1}^{M_a} a_k \, cos(\omega_{a,k} n + \phi_{a,k}) + \sum_{k=1}^{M_b} b_k \, cos(\omega_{b,k} n + \phi_{b,k}) \tag{3.17}$$

or

$$s_c[n] = \sum_{k=1}^{M=M_a+M_b} c_k \, cos(\omega_{c,k} n + \phi_{c,k}) \tag{3.18}$$

Using vector notation, we can simplify (3.18) by rewriting as

$$\boldsymbol{s}_c = \left[\boldsymbol{x}_1^T \cos\left(\boldsymbol{x}_2 \boldsymbol{n}^T + \boldsymbol{x}_3 \boldsymbol{l}^T\right)\right]^T \tag{3.19}$$

where

$$\boldsymbol{x}_1 = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_M \end{bmatrix}_{Mx1} \quad \boldsymbol{x}_2 = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_M \end{bmatrix}_{Mx1} \quad \boldsymbol{x}_3 = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_M \end{bmatrix}_{Mx1} \quad \boldsymbol{n} = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ N \end{bmatrix}_{Nx1} \quad \boldsymbol{l} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{Nx1}$$

Given our co-channel speech data segment $\boldsymbol{s}$, we must select the proper $\boldsymbol{X}$, where

$$\boldsymbol{X} = [\boldsymbol{x}_1 \;\vdots\; \boldsymbol{x}_2 \;\vdots\; \boldsymbol{x}_3] \tag{3.20}$$

such that we minimize the two-norm of the residual vector

$$\boldsymbol{r} = \boldsymbol{s}_c - \boldsymbol{s} \tag{3.21}$$

Thus the function to be minimized becomes the squared error between our model (3.19) and the windowed co-channel speech data $\boldsymbol{s}$. More formally, our problem statement becomes:

Find the proper $\boldsymbol{X}$ to solve:

$$min\, \boldsymbol{F}(\boldsymbol{X}) = \frac{1}{2}[\boldsymbol{s}_c - \boldsymbol{s}]^T \boldsymbol{W}^T [\boldsymbol{s}_c - \boldsymbol{s}] \tag{3.22}$$

where the matrix $\boldsymbol{W}$ is a positive definite diagonal weighting matrix with the diagonal elements equal to coefficients of our window function

$$W = \begin{bmatrix} w[-(N-1)/2] & 0 & 0 & \cdots & \\ 0 & w[1-(N-1)/2] & 0 & \cdots & \\ 0 & 0 & \ddots & 0 & \\ \vdots & \vdots & 0 & w[(N-1)/2] & \end{bmatrix} \tag{3.23}$$

Since $\boldsymbol{W}$ is a matrix with constant values, the proper $\boldsymbol{X}$ that minimizes the residual vector in (3.21) also minimizes the function $\boldsymbol{F}(\boldsymbol{X})$ defined in (3.22).

Let us consider a nonlinear function of the form

$$f(X,n)=x_{11}cos(x_{21}n+x_{31})+x_{12}\,cos(x_{22}n+x_{32})+\ldots+x_{1M}\,cos(x_{2M}n+x_{3M}) \qquad (3.24)$$

or

$$f(X,n)=f_1(x_1,n)+f_2(x_2,n)+\ldots+f_M(x_M,n) \qquad (3.25)$$

where

$$f_j\left(x_j,n\right)=x_{1j}\,cos\left(x_{2j}n+x_{3j}\right) \qquad (3.26)$$

and

$$\boldsymbol{x}_j=\begin{bmatrix} x_{1j} \\ x_{2j} \\ x_{3j} \end{bmatrix} \qquad (3.27)$$

The elements of $\boldsymbol{x}_j$ represent the amplitude, frequency and phase of the *jth* sinusoidal component of *f*(***X***,*n*). Equation (3.24) is our nonlinear function, which is our mathematical representation of overlapping voiced co-channel speech. We want to match this representation to our measured discrete data set, $\boldsymbol{s}$ = $(s_1\ s_2 \ldots s_N)^T$ at each discrete time n_k:

$$\begin{aligned}
&f_1(\boldsymbol{x}_1,n_1)+f_2(\boldsymbol{x}_2,n_1)+\ldots+f_M(\boldsymbol{x}_M,n_1)=s_1\\
&f_1(\boldsymbol{x}_1,n_2)+f_2(\boldsymbol{x}_2,n_2)+\ldots+f_M(\boldsymbol{x}_M,n_2)=s_2\\
&\vdots\\
&f_1(\boldsymbol{x}_1,n_N)+f_2(\boldsymbol{x}_2,n_N)+\ldots+f_M(\boldsymbol{x}_M,n_N)=s_N
\end{aligned} \qquad (3.28)$$

This is an over-determined and consequently inconsistent set of equations where $N > M$ and $N > 3M$. N represents the number of samples in our data segment.

We can define $\boldsymbol{A}$ as our nonlinear function

$$\boldsymbol{A}=\left[A_1 \quad A_2 \quad \cdots \quad A_N\right]^T \qquad (3.29)$$

where

$$A_k = \sum_{i=1}^{M} f_i(\boldsymbol{x}_i, n_k), \qquad \textit{for } k = 1 \textit{ to } N \tag{3.30}$$

Using our nonlinear function, we can express the weighted residual error vector as

$$\boldsymbol{r}_W = \boldsymbol{W}^{1/2}(\boldsymbol{A} - \boldsymbol{s}) = \boldsymbol{W}^{1/2}\boldsymbol{r} \tag{3.31}$$

where the weighting matrix $\boldsymbol{W}^{1/2}$ is defined to be

$$\boldsymbol{W}^{1/2} = \begin{bmatrix} \sqrt{w[-(N-1)/2]} & 0 & 0 & \cdots \\ 0 & \sqrt{w[1-(N-1)/2]} & 0 & \cdots \\ 0 & 0 & \ddots & 0 \\ \vdots & \vdots & 0 & \sqrt{w[(N-1)/2]} \end{bmatrix} \tag{3.32}$$

Each element in $\boldsymbol{r}_W$ represents the weighted residual error at each discrete time between our model and the measured data.

The function to be minimized is the nonlinear least squared objective function, $F_W(\boldsymbol{X},n)$ which is defined to be the two-norm of the weighted residual error

$$F_W(\boldsymbol{X}, n) = \tfrac{1}{2}\boldsymbol{r}_W{}^T\boldsymbol{r}_W = \tfrac{1}{2}\sum_{k=1}^{N}\left\{(r_W)_k\right\}^2 \tag{3.33}$$

The *kth* row of the weighted residual error is the *kth* element of the summation in (3.33) given by

$$(r_W)_k = \sqrt{w_k}\{f_{k1} + f_{k2} + \ldots + f_{kM} - s_k\} \qquad \text{for } k = 1 \text{ to } N \tag{3.34}$$

with the weighting matrix defined as $\boldsymbol{W}^{1/2} = diag\left(\sqrt{w_1} \quad \sqrt{w_2} \quad \cdots \quad \sqrt{w_k} \quad \cdots \quad \sqrt{w_N}\right)$.

The matrix $\boldsymbol{X}^*$ which causes $F_W(\boldsymbol{X})$ to be a minimum must simultaneously solve the gradients to zero [47]

$$\nabla_{lj}F_W = \sum_{k=1}^{N}(r_W)_k\left(\nabla_{lj}(r_W)_k\right) \quad j = 1 \text{ to } M \text{ and } l = 1 \text{ to } 3 \tag{3.35}$$

or

$$\boldsymbol{g_1} = \nabla_{1j}F_W\left(\boldsymbol{X}^*\right) = 0 \tag{3.36}$$

$$\boldsymbol{g}_2 = \nabla_{2j}F_W\left(\boldsymbol{X}^*\right) = 0 \tag{3.37}$$

$$\boldsymbol{g}_3 = \nabla_{3j}F_W\left(\boldsymbol{X}^*\right) = 0 \tag{3.38}$$

for j = 1 to M, while also ensuring that the *Hessian* of the function $F_W(\boldsymbol{X})$, defined as

$$\boldsymbol{H} = \nabla\left(\nabla\boldsymbol{F}(\boldsymbol{X})\right)^T = \nabla^2\boldsymbol{F}(\boldsymbol{X}) \tag{3.39}$$

is positive definite at our solution matrix $\boldsymbol{X}^*$.

The first partial derivative of the *kth* residual error with respect to x_{lj} is

$$\nabla_{lj}(r)_k = \nabla_{lj}f_{k1} + \nabla_{lj}f_{k2} + \ldots + \nabla_{lj}f_{kM} = \sum_{q=1}^{M}\nabla_{lj}f_{kq} \tag{3.40}$$

where k will vary between 1 and N, j will vary from 1 to M and l will vary from 1 to 3. We can now define the *Jacobian matrix* as a matrix of the first partial derivatives of the residual error vector,

$$\boldsymbol{J} = \left[\nabla_{lj}(r)_k\right] \quad \text{for rows 1 to } N \text{ and columns 1 to } 3M. \tag{3.41}$$

Rewriting (3.41) as a weighted *Jacobian matrix* we have

$$\boldsymbol{J}_W = \left[\nabla_{lj}(r_W)_k\right] = \boldsymbol{W}^{1/2}\boldsymbol{J} \tag{3.42}$$

Written out, the *Jacobian* is

$$\boldsymbol{J}_W = \begin{bmatrix} \nabla_{11}(r_W)_1 & \nabla_{12}(r_W)_1 & \nabla_{13}(r_W)_1 & \cdots & \nabla_{M1}(r_W)_1 & \nabla_{M21}(r_W)_1 & \nabla_{M3}(r_W)_1 \\ \nabla_{11}(r_W)_2 & \nabla_{12}(r_W)_2 & \nabla_{13}(r_W)_2 & \cdots & \nabla_{M1}(r_W)_2 & \nabla_{M2}(r_W)_2 & \nabla_{M3}(r_W)_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \nabla_{11}(r_W)_N & \nabla_{12}(r_W)_N & \nabla_{13}(r_W)_N & \cdots & \nabla_{M1}(r_W)_N & \nabla_{M2}(r_W)_N & \nabla_{M3}(r_W)_N \end{bmatrix}_{Nx3M} \tag{3.43}$$

The *Jacobian* simplifies the expression in (3.35) to

$$\nabla \boldsymbol{F}_W = \boldsymbol{g}_W = \boldsymbol{J}_W^T \boldsymbol{r}_W \tag{3.44}$$

To find the minimum of our multidimensional objective function $\boldsymbol{F}_w(\boldsymbol{X},n)$,we must find the $\boldsymbol{X}^*$ such that gradient of the objective function is equal to zero and the *Hessian* of our function is positive-definite at the point. The *Hessian* matrix $\boldsymbol{H}$ is composed of all second derivatives of $\boldsymbol{F}(\boldsymbol{X})$. This can be written as

$$\boldsymbol{H}_W = \nabla(\nabla \boldsymbol{F}_W)^T = \nabla^2 \boldsymbol{F}_W \tag{3.45}$$

where$\nabla^2 \boldsymbol{F}_W$ is a matrix of second partial derivatives of $\boldsymbol{F}$. We can substitute (3.35) into (3.45) to obtain

$$\boldsymbol{H}_W = \nabla \left[\sum_{k=1}^{m} (\boldsymbol{r}_W)_k \nabla (\boldsymbol{r}_W)_k \right]^T \tag{3.46}$$

Simplifying (3.46) we obtain

$$\boldsymbol{H}_W = \sum_{k=1}^{N} \nabla \left\{ (\boldsymbol{r}_W)_k (\nabla (\boldsymbol{r}_W)_k)^T \right] \tag{3.47}$$

$$\boldsymbol{H}_W = \sum_{k=1}^{N} \left\{ \nabla (\boldsymbol{r}_W)_k \nabla (\boldsymbol{r}_W)_k^T + (\boldsymbol{r}_W)_k \nabla^2 (\boldsymbol{r}_W)_k \right\} \tag{3.48}$$

$$\boldsymbol{H}_W = \boldsymbol{J}_W^T \boldsymbol{J}_W + \boldsymbol{M}_W \tag{3.49}$$

where $\boldsymbol{M}_w$ is an $3M$x$3M$ matrix of the weighted residuals and their second derivatives defined to be

$$M_W = \sum_{k=1}^{N} (r_W)_k \nabla^2 (r_W)_k \tag{3.50}$$

When the residuals approach zero, near a solution X^*, then equation (3.49) can be approximated by the first term $J_W^T J_W$, since the second term M_w is close to zero when the $(r_w)_k$ are close to zero [49]. If this situation occurs, then (3.49) can be rewritten as

$$H_W = J_W^T J_W \tag{3.51}$$

which is equivalent to assuming that the residuals in (3.31) are linear.

We have developed the proper relationships for our model to accurately estimate our measured data ensuring the least squared error. The optimal solution X^* is obtained when it forces (3.44) to zero while maintaining (3.49) to be positive-definite. We now present an iterative technique to obtaining the nonlinear least squared error.

Let us begin by first recalling that an infinite Taylor series expansion about a particular point x_0 is given by [50]

$$f(dx) = f(x_0) + f'(x_0)dx + \left(\tfrac{1}{2!}\right)f''(x_0)dx^2 + \left(\tfrac{1}{3!}\right)f'''(x_0)dx^3 + \ldots \tag{3.52}$$

where dx represent the distance or displacement from our expansion point x_0

$$dx = (x - x_0) \tag{3.53}$$

From (3.52) we can write the multivariable Taylor series expansion for a function about a vector $\boldsymbol{p}$ as

$$\boldsymbol{F}(\boldsymbol{dx}) = \boldsymbol{F}(\boldsymbol{p}) + \boldsymbol{g}(\boldsymbol{p})^T \boldsymbol{dx} + \tfrac{1}{2}\boldsymbol{dx}^T \boldsymbol{H}(\boldsymbol{p})\boldsymbol{dx} + \ldots \tag{3.54}$$

where the multidimensional displacement vector $\boldsymbol{dx}$ is given as

$$\boldsymbol{dx} = (\boldsymbol{x} - \boldsymbol{p}) = (dx_1 \quad dx_2 \quad \cdots \quad dx_M)^T \tag{3.55}$$

The gradient vector of $\boldsymbol{F(X)}$ evaluated at $\boldsymbol{p}$ is $\boldsymbol{g(p)}$ and the *Hessian* matrix of partial derivatives evaluated at $\boldsymbol{p}$ is $\boldsymbol{H(p)}$. We can approximate our multidimensional nonlinear objective function as a quadratic function (of the same dimension) using the Taylor series expansion given in (3.54). If $\boldsymbol{F(X)}$ quadratic, then

$$\nabla \boldsymbol{F}(\boldsymbol{dx}) = \nabla\left(\boldsymbol{F}(\boldsymbol{p}) + \boldsymbol{g}(\boldsymbol{p})^{T}\boldsymbol{dx} + \tfrac{1}{2}\boldsymbol{dx}^{T}\boldsymbol{H}(\boldsymbol{p})\boldsymbol{dx}\right) \tag{3.56}$$

or

$$\nabla \boldsymbol{F}(\boldsymbol{dx}) = \boldsymbol{g}(\boldsymbol{p}) + \boldsymbol{H}(\boldsymbol{p})\boldsymbol{dx} \tag{3.57}$$

The proper step (magnitude and direction) that must be taken to obtain the minimum would be the $\boldsymbol{dx}$ that solves $\nabla \boldsymbol{F}(\boldsymbol{dx}) = 0$ from any point $\boldsymbol{p}$ on a quadratic surface is

$$\boldsymbol{dx}^{*} = -\boldsymbol{H}(\boldsymbol{p})^{-1}\boldsymbol{g}(\boldsymbol{p}) \tag{3.58}$$

This is known as the *Newton step* in the *Newton-Raphson* search procedure with $\boldsymbol{x} = \boldsymbol{p} + \boldsymbol{dx}^{*}$ called the *Newton point* [47]. The convergence of $\boldsymbol{x}$ to $\boldsymbol{x}^{*}$ such that (3.57) is zero becomes an iterative procedure. If the initial guess of the starting point in estimating the solution to our nonlinear objective function is close to the correct solution, then our Taylor series approximation, given in equation (3.56), will represent the function reasonably well and convergence to a solution is expected.

Referring back to our nonlinear least squared error objective function, the search direction to obtain a solution vector $\boldsymbol{X}^{*}$ that minimizes (3.33) is given by (3.58). Since (3.33) is not quadratic, we can estimate it as a quadratic using the Taylor series expansion given in (3.54). Then a minimum solution is obtained by performing a sequence of steps in the magnitude and direction given by (3.58), calculating the gradient and estimating the *Hessian* at each step. Substituting (3.49) into (3.58) we obtain our *Newton step* as

$$\boldsymbol{dX} = -\left(\boldsymbol{J}_W^T\boldsymbol{J}_W + \boldsymbol{M}_W\right)^{-1}\boldsymbol{J}^T\boldsymbol{r}_W \tag{3.59}$$

We can rewrite (3.59) in terms of the residual error vector and the weighting matrix by substituting (3.31) and (3.42) into (3.59) and assuming the residual error vectors are linear ($\boldsymbol{M} = 0$), we then have

$$\boldsymbol{dX} = -\left(\boldsymbol{J}^T\boldsymbol{W}\boldsymbol{J}\right)^{-1}\boldsymbol{J}^T\boldsymbol{W}\boldsymbol{r} \tag{3.60}$$

Back to our problem of separating overlapping voiced speech, we can replace our nonlinear function in (3.24) with our co-channel voiced speech model in (3.19). The solution to (3.22) then becomes an iterative process of calculating the gradient of the nonlinear least squared objective function defined in (3.22) using equation (3.35) and estimating the *Hessian* in (3.50) at each iterative step based on the direction and magnitude of (3.60).

The search for a solution vector $\boldsymbol{X}^*$ can be obtained with less error and with less iteration when we impose restrictions on the unknown variables. The method of establishing upper and lower bounds on some or all the variables in $\boldsymbol{X}$ is referred to as a box constraint. Less iteration will be required because the constraint imposes limitations on the step size and appropriate regions for convergence. This also allows us to express the *Hessian* using (3.51). The box constraint can be represented as a vector defining the upper bound, $\boldsymbol{X}_{UB}$ and a vector defining the lower bound $\boldsymbol{X}_{LB}$ such that we impose $\boldsymbol{X}$ to be restricted to the region bounded by these vectors. Formally our constraint becomes:

Constraint: The solution vector X^* must lie in the region

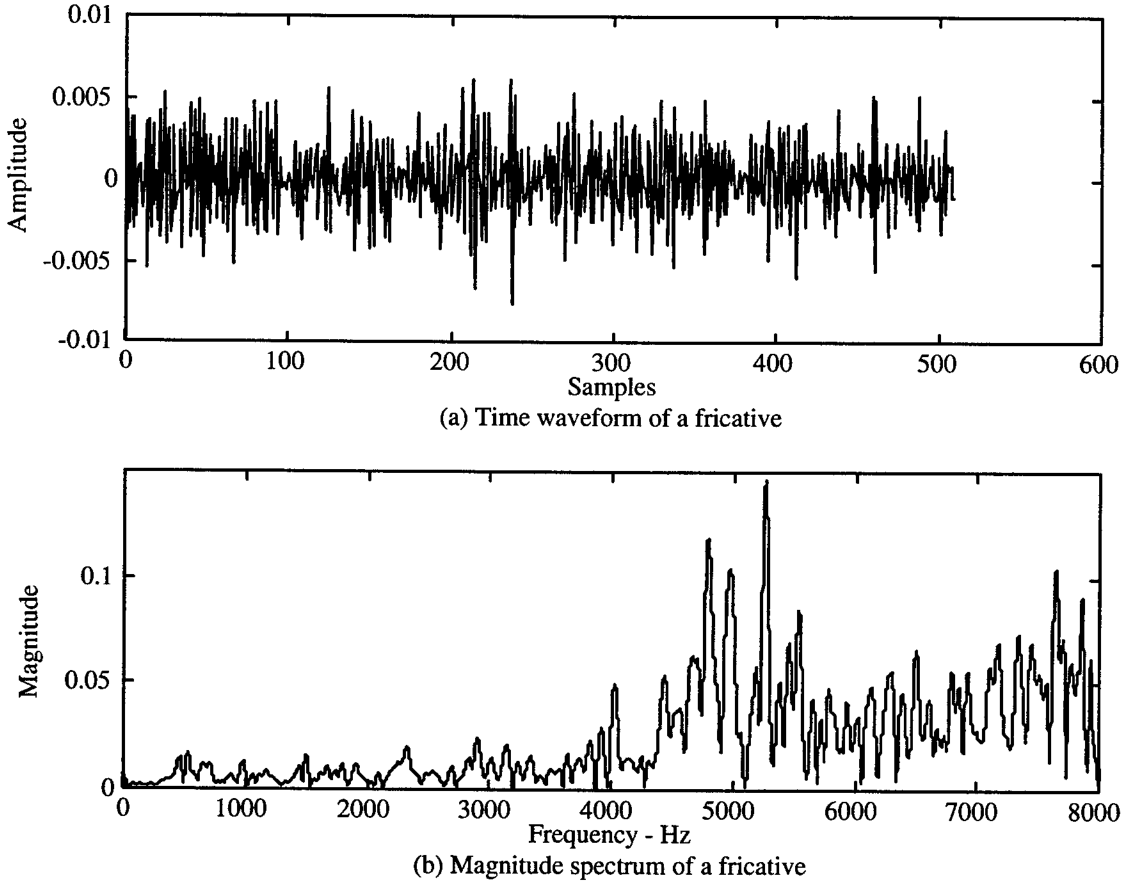
$$X_{LB} \leq X^* \leq X_{UB} \tag{3.61}$$

3.3 Adaptive Filtering (V/UV, UV/V, and UV/UV)

We have just developed a technique to separate two overlapping voiced speech signals. Referring to Figure 3.9, our next step is to separate voiced speech from unvoiced speech and overlapping unvoiced speech. Separating overlapping speech signals in which one signal is voiced and the other is unvoiced can be a difficult problem due to the lack of a parametric model for mixed speech, such as the sinusoidal model that exists for overlapping voiced speech.

Two examples of unvoiced sounds are a fricative and a plosive. Examples of these two speech sounds are given in Figure 3.10 and Figure 3.11 respectively. A frication is the result of air flowing past a constriction in the vocal tract, which generates a broadband noise sound. A plosive or stop is the result of a sudden release of air pressure that has been built up from a closure in the vocal tract.

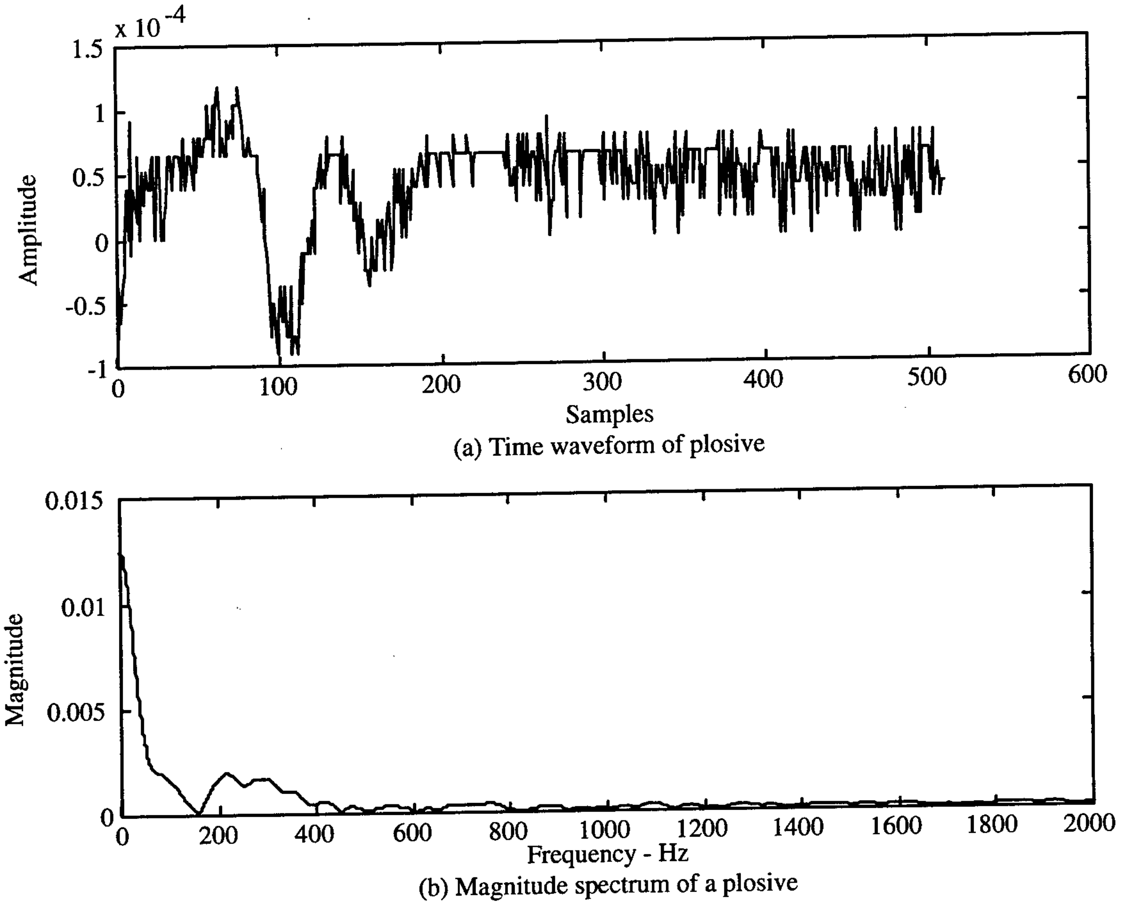
In Figure 3.10, it can be seen that for most types of unvoiced speech sounds (such as fricatives) a major portion of their energy is concentrated at higher frequencies (above 4 kHz). We have implemented a lowpass to separate an interfering unvoiced signal from the desired voiced signal.



(a) Time waveform of a fricative

(b) Magnitude spectrum of a fricative

Figure 3.10: Representation of a Fricative, (a) time waveform and (b) magnitude spectrum.



(a) Time waveform of plosive

(b) Magnitude spectrum of a plosive

Figure 3.11: Representation of a Plosive, (a) time waveform and (b) magnitude spectrum.

When an interfering signal is voiced and the desired signal is unvoiced the energy of the interfering signal is concentrated at lower frequencies and the energy of the desired signal is concentrated at the higher frequencies. We have implemented a highpass filter to separate the voiced interference signal from the unvoiced desired signal.

Referring back to the tests we conducted in Section 3.1, it was found that the intelligibility of speech, using a lowpass/highpass filter to separate overlapping voiced speech sounds from unvoiced speech sounds, is high.

Cherry and Wiley [45] reported that speech from which non-vocalic (unvoiced) sounds had been gated out and removed, regained their intelligibility significantly when wideband noise was inserted in those intervals. Apparently, the brain will accept any suitably placed noise burst as the required sound. Parsons [3] also noted that when only the periodic portions of speech are played back, the intelligibility of the speech is still clear and realistic. Therefore it is necessary to only attenuate overlapping unvoiced co-channel speech segments (unvoiced/unvoiced), consistent with the energy level in the reconstructed speech signal.

A concern, stated in Chapter 2 was the assumption that the speech signals must be stationary during the windowed interval. This may not always be the case. We have implemented a discriminant to estimate nonstationarity within a given windowed segment of speech. This discriminant is the ratio of the short-time energy measure in the first half of the segment to the short-time energy measure in the second half of the segment, defined in equation (2.23). When this threshold is exceeded, the window length is halved and processing is conducted on a time interval half the original interval time. This is a rather simple, but effective approach to insuring stationarity during the analysis interval.

3.4 Voicing State Determination

In this section, we present a voicing state determination algorithm (VSDA) to estimate the voicing state of a segment of co-channel speech. Voicing state determination is a method of classifying the voicing state of the speakers present in a segment of co-channel speech. This process is required in a co-channel speaker separation system as a means to select an appropriate separation processing technique. The possible voicing state classifications are:

1. Silence (S) - both speakers are silent;
2. Voiced/Voiced (V/V) - both speakers are producing voiced sounds;
3. Voiced/Unvoiced (V/UV) - the desired speaker is producing voiced sounds while the interfering speaker is producing unvoiced sounds;
4. Unvoiced/Voiced (UV/V) - the desired speaker is producing unvoiced sounds and the interfering speaker is producing voiced sounds;
5. Unvoiced/Unvoiced (UV/UV) - both speakers are producing unvoiced sounds.

Classifying co-channel speech requires simultaneously estimating the voicing state of each speaker present within the segment. We have assumed the silent state as a subset of the unvoiced class (except when both speakers are silent) thereby limiting classification of co-channel speech to mixtures of voiced and unvoiced speech and total silence.

In this work we have developed a technique using three different classifiers to perform voicing state determination based on decision theory. Our detector can be modeled as a black box with a set of inputs and a set of outputs. The box operates in both a training mode and a detection mode. In training mode, the detector is presented with co-channel speech data segments from which it then creates a reference associated with the five classes defined above. Once training is complete, the detector operates in a recognition mode in which it is presented with an unknown set of data. The detector is then tasked to identify which of the five possible voicing classes should be assigned to the data. The detector is evaluated based on its ability to correctly classify unknown co-channel speech segments. A Bayesian classifier, a *k-nearest neighbor* classifier, and a Parzen window classifier are developed below. Each classifier used the same decision structure and the same feature set to classify speech.

3.4.1 Decision Structure

Voicing state determination of co-channel speech requires discrimination between five classes or categories of speech. There are several ways in which an R-category (R = 5 case) classification can be structured. Classification can be obtained using a single classifier that assigns the pattern to one of R classes, or through a sequence of binary decisions. We have chosen the binary decision tree approach to classification.

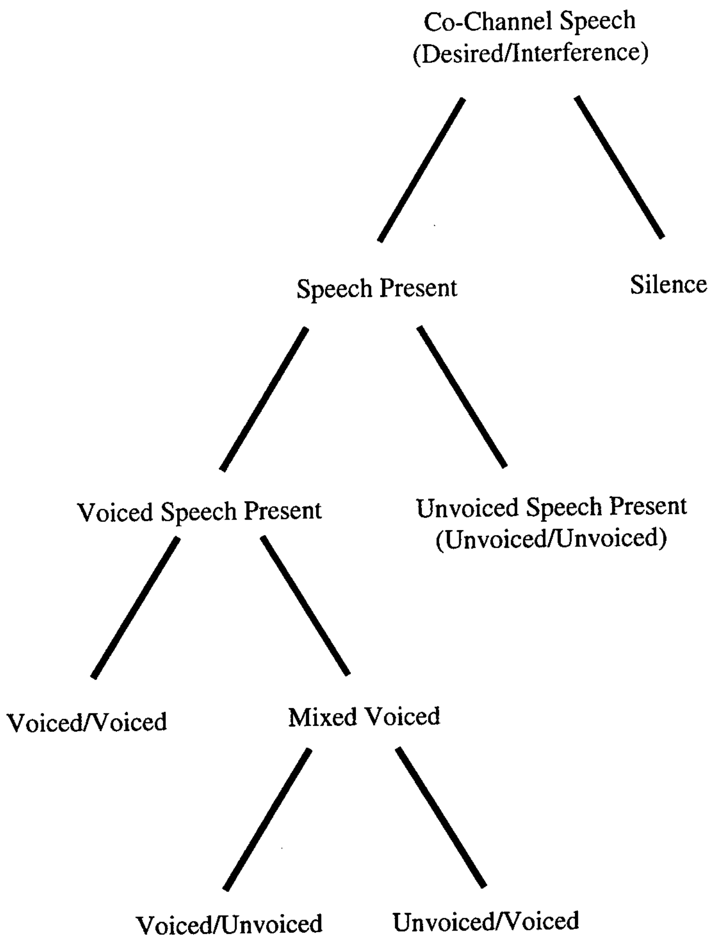


Figure 3.12: Voicing state decision tree for co-channel speech.

Our binary decision tree structure is shown Figure 3.12. Decisions are made independently, on a frame-by-frame basis. The first decision is to decide on the presence or absence of speech in the given speech segment. If the decision is made that no speech is present, then the segment is labeled as *silence*. If speech is present, we move down the decision tree to the next level. Here a decision must be made on the presence of any voiced speech or the presence of strictly unvoiced speech. If only unvoiced speech is present, the segment of speech is labeled as *unvoiced/unvoiced.* If voiced speech is present, we proceed down the decision tree to decide if both speech segments contain voiced speech or if one sound is voiced and the other is unvoiced. If both speech sounds are voiced, we label the speech as *voiced/voiced.* If the speech segment is mixture of voiced and unvoiced speech, we continue down to the last branch to decide which speaker is voiced and which is unvoiced. Here the speech is labeled as *voiced/unvoiced* or *unvoiced/voiced.* Determination as to the presence or absence of voicing is used as a means to select the appropriate separation processing technique.

3.4.2 Features

The selection of a set of features that will provide adequate classification of co-channel speech must be more sophisticated than those used for voicing state classification of uncorrupted speech. The set of features chosen must not only discriminate between classes of voiced and unvoiced speech, but it must also discriminate between mixed excitation of two speakers. That is, the feature set must successfully discriminate between the sum of two voiced segments of speech from the sum of a voiced and unvoiced speech segment. The feature set must also discriminate between mixed excitation between two different speakers. The features we have chosen are:

1. Log of the short time energy of the signal (STE);
2. Normalized fundamental frequency (PIT);
3. Normalized autocorrelation coefficient at unit sample delay (MAXAC);

4. Normalized zero crossing rate (ZCR);

5. Ratio of energy in the signal above 4 kHz to energy below 4 kHz (HILO);

6. 16 mel-cepstral coefficients (MELCEP);

7. 15 modified covariance coefficients, excluding the first coefficient (MCV).

The features considered here are chosen not only for their ability to discriminate between voiced, unvoiced and mixed speech, but also to differentiate between speakers. The first four features are a subset of the traditional voicing state determination systems. The last two features in the set are unique to our application in the discrimination of voiced/voiced speech from mixed voiced speech and in discriminating mixed voiced speech between speakers.

The STE of the speech signal is calculated using equation (2.23). The PIT feature is measured based on the technique presented in Section 3.5. The value of MAXAC is taken as the ratio of the maximum value over the difference between the maximum and minimum value of equation (2.22). The ZCR is the total number of zero crossings as defined in equation (2.30). The value of HILO is dependent on the spectral characteristics of the speech segment. The speech segment is low-passed filtered with a cut-off of 4 kHz and equation (2.23) is used to calculate the energy. The speech segment is then high-passed filtered with a cut-off of 4 kHz and again equation (2.23) is used to calculate the energy. The value of HILO is the ratio of these two values. The Mel-cepstral coefficients are calculated using equation (2.32). The number of frequency bins corresponds to a center frequency spacing of 150 mels. The 16 mel-frequency weighted cepstral coefficients does not include the zero'th order coefficient. An algorithm for calculating the modified covariance coefficients can be found in Marple [41]. Here we have chosen the MCV coefficients immediately following, but not including the zero'th order coefficient.

3.4.3 Training Data

In this research we are developing a pattern recognition approach for deciding voicing state of speech based on measured features from the co-channel speech signal. Our classifier is trained to recognize patterns of speech through supervised learning. Training is performed on the uncorrupted speech of each speaker in the co-channel signal. We have conducted training using the TIMIT database. A more detailed description of our data can be found in Chapter 4. The TIMIT database contains clean English spoken speech sampled at 16 kHz. The database is segmented into eight distinct dialect regions of the United States. These regions include New England, Northern USA, North Midland USA, South Midland USA, Southern USA, New York City, Western USA and Army Brat (an individual who has moved around). We have performed training and testing on Northern USA speakers.

The TIMIT database provides a hand labeled phonetic transcription of each sentence within the database. Logically, this would appear to be the most accurate way to segment the speech. However, since a typical phone will transverse across several frames, and a phone could contain both voiced and unvoiced speech, we have developed our own segmentation and labeling system for training. This will also aid in training and testing using the RPI_COC speech database, which is unlabeled data. The features, described above, are extracted from the labeled data and used to train the classifier.

The technique used to segment uncorrupted speech for training our classifier is shown in Figure 3.13. The short time energy, defined in (2.23) along with the zero crossing rate (2.30) are two features that have proven to be effective in making a voiced/unvoiced classification of uncorrupted speech [34].

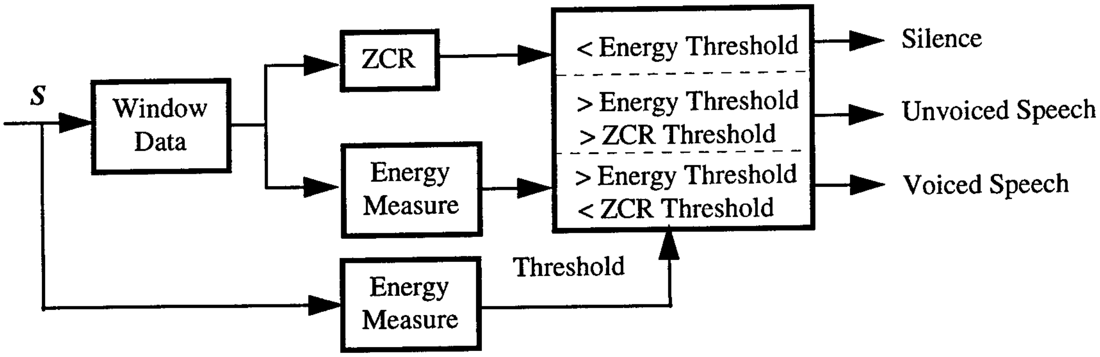


Figure 3.13: Voiced/Unvoiced segmentation of uncorrupted speech.

We define the energy threshold as

$$E_{thrshld} = .6 * min\left(\frac{1}{M} \sum_{i=1}^{M} 10 * log(E_i) \right) \qquad (3.62)$$

where E_i is the short time energy measure per frame and M is the total number of frames within the length of the co-channel speech signal. The energy threshold is used to determine periods of silence.

The threshold for zero-crossing rate (ZCR) is predetermined based on the sampling rate and the frame size used in the windowing routine. The threshold for the ZCR is given by [40]

$$ZCR_{Thrhld} = \frac{2480}{f_s} * N \tag{3.63}$$

where f_s is the sampling rate and N is the number of samples per frame. When the ZCR is greater than the given threshold and the energy is greater than the threshold, the segment of speech is labeled as unvoiced. If the ZCR is less than or equal to the threshold and the energy is greater than the threshold, the speech segment is labeled as voiced. Otherwise the energy within the speech segment is below the threshold and the speech is labeled silence.

3.4.4 Bayesian Classifier

Based on the decision structure presented above, our problem of identifying the voicing states of speech segments becomes a sequential series of decisions between two classes. Therefore, in the following sections, we treat the classification of speech as a two-class problem. In our two-class problem, hypothesis H_0 is true when X belongs to class 0 and hypothesis H_1 is true when X belongs to class 1.

The Bayes decision rule for minimum error, on a two-class problem is; given an observation vector $\boldsymbol{x}$, the classifier decides hypothesis H_0 if the probability of H_0 is greater than the probability of H_1. Otherwise, the decision is H_1. This can be written as a likelihood ratio test [51]

$$\ell(x) = \frac{p(x/H_1)}{p(x/H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_0}{P_1} \qquad (3.64)$$

where the P_i is the *a priori* probability of hypothesis H_i , and $p(x/H_i)$ is the conditional density function. The term $\frac{P_0}{P_1}$ becomes the threshold value of the likelihood ratio decision. We assume there is no cost associated with a correct decision and the costs associated with a wrong decision are equal.

To achieve minimum error rate classification under the Bayes decision rule, we must chose our classification such that it minimizes the conditional risk. Thus, we must decide the hypothesis that maximizes the a posteriori probability $p(H_i/x)$. The hypothesis with the largest a posteriori probability insures a minimum error rate.

The form of the classifier is dependent on the conditional density functions $p(x/H_i)$. The likelihood ratio takes on an analytically attractive form when the density functions are multivariate normal. A multivariate normal density function is defined as

$$p(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right] \qquad (3.65)$$

where μ is the *n*-component mean vector and Σ is the *n-by-n* covariance matrix. Unfortunately, our *n*-dimensional vector is not multivariate normal. However, we can form a linear combination of the components of x that will project this *n*-dimensional vector onto a line. We can write this projection as

$$y = w^t x \qquad (3.66)$$

where y is a linear sum of the elements of **x**. If this transformation is chosen properly, we can project these vectors in such a manner that the samples are well separated.

To insure that the samples are well separated, the distance between the means of the projected samples must be large while maintaining the variances of these projected samples to be small. The Fisher linear discriminant [52] is a linear function $\boldsymbol{w}$ defined in equation (3.66) such that the criterion function

$$J(\boldsymbol{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \tag{3.67}$$

is maximum, where $\tilde{m}_i$ is the sample mean of the projected points for class i, and $\tilde{s}_i$ is the scatter of the projected samples for class i. Then $\boldsymbol{w}$ is given by

$$\boldsymbol{w} = S_W^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{3.68}$$

where $\boldsymbol{m}_i$ is the n-dimensional sample mean of the points for class i, and S_w is the *within-class scatter matrix*

$$S_W = S_1 + S_2 \tag{3.69}$$

with

$$S_i = \sum_{\boldsymbol{x} \in R_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^t \tag{3.70}$$

If the elements of $\boldsymbol{x}$ are mutually independent, the dimension of $\boldsymbol{x}$ is large and the $w_i X_i$ satisfies the Lindeberg conditions, then from the *central limit theorem*, y can be taken to be a normal random variable. The Lindeberg condition states that the individual variances σ_k^2, for k = 1,...,n must be small compared to the sum of all the variances,

$\sum_{i=1}^{n}\sigma_i^2$. The assumption that y is a normal random variable provides an optimum partitioning of the real line into two decision regions.

Referring back to equation (3.64), we must now develop a classifier based on the statistical characteristics of our sampled data. We can view the likelihood ratio test in (3.64) in terms of a set of discriminant functions $g_i(\boldsymbol{x})$ for each hypothesis or class. Our classifier assigns an observation $\boldsymbol{x}$ to the class with the largest discriminant. For the minimum error rate, our discriminant functions becomes

$$g_i(\boldsymbol{x}) = p(H_i/\boldsymbol{x}) \tag{3.71}$$

such that the maximum discriminant function is the maximum a posteriori probability. Using the Bayes rule, we can rewrite (3.68) as

$$g_i(\boldsymbol{x}) = p(\boldsymbol{x}/H_i)P(H_i) \tag{3.72}$$

By taking the logarithm of both sides, the classification will not change since the logarithm is a monotonically increasing function. Then (3.72) becomes

$$g_i(\boldsymbol{x}) = log\, p(\boldsymbol{x}/H_i) + log\, P(H_i) \tag{3.73}$$

By substituting equation (3.65) into (3.73), we can express our discriminant function as

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^t\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{n}{2}log(2\pi) - \frac{1}{2}log|\Sigma_i| + log\, P(H_i) \tag{3.74}$$

This expression can be further simplified by applying the linear transformation obtained from (3.68) in which the multivariate normal density function is transformed to a univariate normal density function provided the conditions of the central limit theorem hold. Assuming that the variances for each class are not equal, then (3.74) can be written as

$$g_i(y) = -\frac{1}{2}(y - m_i)^t \sigma_i^{-1}(y - m_i) - \frac{1}{2} log(|\sigma_i|) + log\, P(H_i) \qquad (3.75)$$

The constant term $\frac{n}{2} log(2\pi)$ term has been dropped since it is independent of the voicing state of the signal. The discriminant functions are quadratic and the decision regions lie along a straight line.

This procedure is not optimal. Projection of an *n-dimensional* vector onto a real line can not reduce the minimum achievable error rate. We are throwing away information that may aid in the classification. Also, while the assumption that y is a normal random variable will not always be true, the Bayes classifier may still provide an optimal partitioning of the real line. This technique does allow us to use the Bayes rule applied to a normal density function, which is mathematically attractive and has the added advantage of working in a single dimension.

Next we investigate two non-parametric classifiers which theoretically will give error rates which are greater than the Bayes rate but does not assume underlying normal statistics.

3.4.5 k-Nearest Neighbor Classifier

The nearest neighbor rule for classifying an *n-dimensional* observation feature vector $\boldsymbol{x}$ is, given a set of samples $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, the classifier assigns our observation vector to the class associated with the sample vector, $\boldsymbol{x}_{nn}$, nearest to $\boldsymbol{x}$ [52]. As with the above procedure, the nearest-neighbor rule is a sub-optimal technique. Given a large set of samples, the nearest-neighbor rule tends to work well based on the assumption that the a posteriori probabilities are equal

$$P(H_i/\boldsymbol{x}_{nn}) \approx P(H_i/\boldsymbol{x}) \tag{3.76}$$

since $\boldsymbol{x}_{nn}$ is sufficiently close to $\boldsymbol{x}$.

We can extend the nearest-neighbor rule to the k-nearest-neighbor rule. This rule follows along the same lines as the nearest-neighbor rule except that the observation vector is assigned to the class with the greatest frequency among the k nearest neighbors (*kNN*). Typically this is accomplished using a voting method among an odd-number of nearest neighbors. The observation vector is assigned to the class represented by the majority of the *kNN*'s.

As the value of k increases, the error rate of the *k-nearest-neighbor* rule approaches the Bayes minimum error rate. Therefore, a large value for k will provide a reliable estimate. However, the value of k is limited by the dimension of the observation vector and the total number of vectors in the sample set. Also, the cost associated with a large k is in time and processing. When the dimension of the feature vector is large ($n>>1$) and the number of samples is also large, the search algorithm can be quite extensive. There are clustering methods that have been developed, such as the branch and bound method [53] that can be applied to the sample space to reduce the search time.

We can also apply the Fisher linear discriminant (3.67) to the samples, as described above, to project the data into the best single dimension.

3.4.6 Parzen Window Classifier

The Parzen density estimate and the *k-nearest neighbor* density estimate are fundamentally very similar, but exhibit some different statistical properties. In the *kNN* approach, we fixed *k* (the number of sample vectors closest to our observation vector) and let the local region, *v*, around our observation vector, which contained the *k* samples to be a random variable. In the Parzen density estimate, we fix the local region, *v*, and let ***k*** be a random variable.

If we define the local region around our observation vector to be *L(X)*, then the probability mass of *L(X)* can be approximated by *p(X)v*, where *v* is the volume of *L(X)*. Given a large number of samples, *N*, drawn from *p(X)*, the probability mass can be estimated by counting the number of samples, *k*, within *L(X)* and computing *k/N*. We can then estimate the density function as

$$\hat{p}(X) = \frac{k}{Nv} \tag{3.77}$$

We can also set up a kernel function, $\kappa(X)$, with a volume *v* and height 1/*v*, around all existing samples. Then the average of the values of these kernel functions at *X* is

$$\hat{p}(X) = \frac{1}{N}\sum_{i=1}^{N}\kappa(X - X_i) \tag{3.78}$$

where the shape of the kernel function can take on any complex shape, provided $\int \kappa(X)dX = 1$. Typical kernel functions are either a normal or uniform kernel.

Once we have an estimate of our probability mass function, we can apply the likelihood ratio classifier similar to the one given in (3.64). Then our likelihood ratio test becomes

$$\ell(x) = \frac{\hat{p}_1(X)}{\hat{p}_0(X)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\mathrm{P}_0}{\mathrm{P}_1} \tag{3.79}$$

where we have determined the threshold based on the Bayes criterion. The dimensions of this data can also be reduced using the Fisher linear discriminant function (3.68).

3.5 Joint Pitch Estimation

Pitch is a measure of the fundamental frequency of voiced speech. Joint pitch estimation is a process by which an estimate is made of all fundamental frequencies present in a segment of co-channel speech. The number of fundamental frequencies is directly related to the number of voiced speech signals present in the co-channel signal. The pitch estimate is a crucial parameter for separating co-channel speech. Most separation systems rely on *a priori* pitch contours obtained from the uncorrupted speech signals prior to mixing. In real world applications, this information is not available.

We have investigated and tested several techniques to estimate the pitch period of overlapping voiced speech. Based on these results, we have developed a technique to estimate the pitch contour of both signals.

3.5.1 Joint Pitch Estimation Techniques

Peak-related methods offer the most promise in estimating the pitch from an uncorrupted speech signal. These techniques can be modified to measure the pitch frequencies of co-channel speech. Very few methods employ a single look approach in which both frequencies are measured simultaneously. Most methods are iterative. A measure of the stronger pitch frequency is made and then used to suppress the stronger harmonics. An estimate of the pitch of the weaker signal is then made from this residual signal. This technique depends on accurate measurement of the stronger pitch frequency. We present an investigation of three techniques. The modified covariance method is used to estimate the pitch frequencies simultaneously (single look). The modified autocorrelation and the maximum likelihood pitch estimators are evaluated, based on an iterative approach, using co-channel speech.

The modified covariance (MCV) estimator is based on an autoregressive model of the vocal tract filter in which the linear prediction sequence is solved by minimizing the forward and backward prediction squared errors, described in Marple [41]. In work by Naylor and Porter [24] it was found that the magnitude spectrum of a (zero padded) sequence of the MCV coefficients reveals sharp spectral peaks (when inverted) at the location of narrow sinusoidal components. By first lowpass filtering the data then downsampling and applying a DFT to the MCV coefficients, the harmonics of both the stronger speaker and the weaker speaker become visible. A sophisticated clustering algorithm is used to group the spectral peaks, which are harmonically related, to reasonable pitch values. An example of the spectrum of the MCV coefficients of overlapping vocalic speech at SIR of 0 and -6 dB is provided in Figure 3.14. The segment of overlapping vocalic speech has fundamental frequencies of 111 Hz and 176 Hz respectively. The MCV analysis estimated the pitch frequencies to be 99 and 172 Hz at SIR = 0 dB and 128 and 176 Hz at SIR = -6 dB.

While the MCV estimator produced estimates for two pitch frequencies, the technique was susceptible to pitch doubling errors, frequency shifts, and errors due to fluctuations in SIR. The system is not reliable due to the requirement of a sophisticated clustering routine that must be able to differentiate between a pitch peak and a pitch doubling peak associated with a stronger or weaker speaker.

The autocorrelation technique has been used quite extensively for pitch estimation of uncorrupted speech [43]. We have modified this technique by performing a nonlinear transformation (data cubing) of the speech signal prior to pitch estimation. This nonlinear processing attenuates the lower frequencies and tends to force the pitch estimator to measure the higher pitch frequency. A block diagram of this technique is given in Figure 3.15. The data cubing enhances the periodicity of the stronger speech signal by suppressing the low amplitude portions (i.e. weaker speech signal) of the speech signal. This will enhance the F_0 formant of the stronger signal and suppress the F_0 format of the weaker speech signal. This cubing operation should also reduce the effects of pitch doubling, a common problem in pitch determination in which the higher formants (low amplitude portion of a speech segment) are erroneously labeled as the F_0 formant.

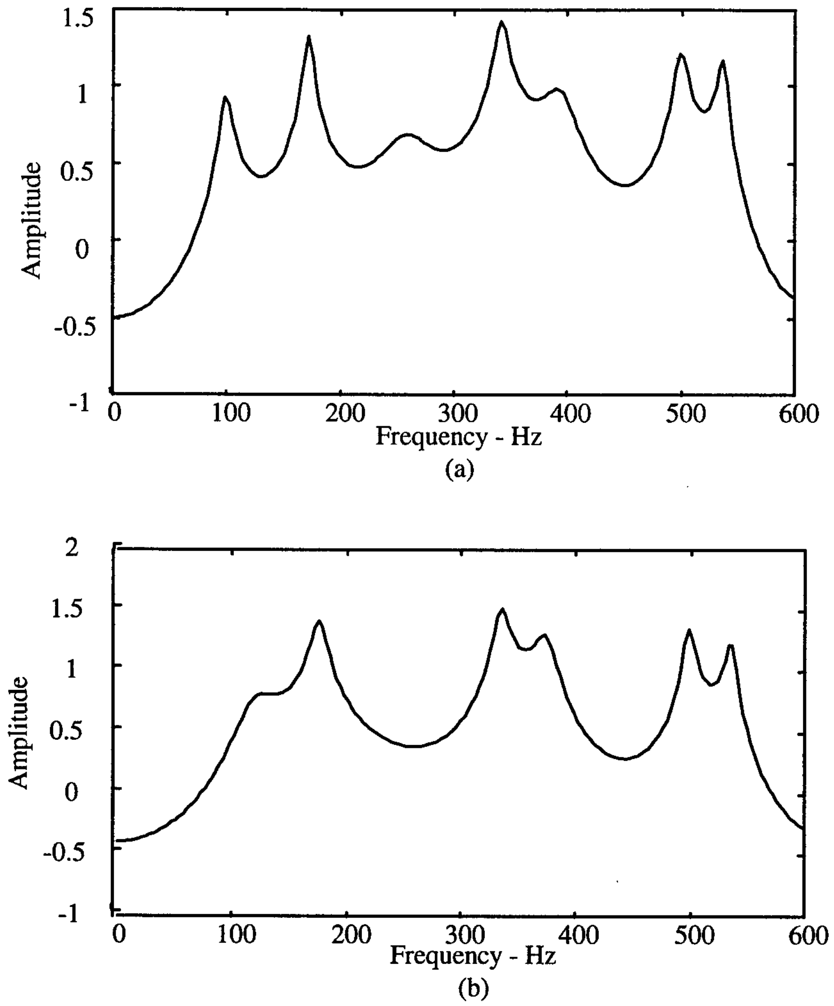


Figure 3.14: Spectrum of MCV coefficients of overlapping vocalic speech segment: (a) SIR = 0 dB, pitch frequencies measured at 99 Hz and 172 Hz, and (b) SIR = -6 dB, pitch frequencies measured at 128 Hz and 176 Hz. True pitch frequencies are 111 and 176 Hz.

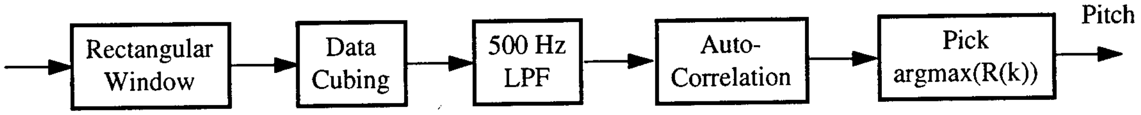


Figure 3.15: Modified autocorrelation pitch estimation technique using nonlinear preprocessing.

The iterative approach is highly dependent on the ability of the pitch estimator to accurately measure the fundamental frequency of the stronger speaker in the presence of another speech signal. An example of the modified autocorrelation technique's ability to measure the pitch of the stronger speaker is presented in Figure 3.16. In this example, it erroneously measured the pitch at 113 Hz. The true measurement should be 176 Hz. The error occurred because the F_0 formant of the weaker signal constructively added with a weaker formant of the stronger speaker. The autocorrelation method only considers the lag associated with the F_0 formant, which makes this routine more susceptible to pitch errors when the higher formant frequencies overlap.

The maximum likelihood estimation (MLE) method measures the pitch estimate based on the contribution of all the formants within a signal [30]. By considering the contribution of all the formants, the MLE method is less susceptible to pitch doubling errors than the modified autocorrelation method. A block diagram of the MLE method is given in Figure 3.17. In the frequency domain, this periodic estimator can be interpreted as the inner product of a comb filter (inner spacing P) with the autocorrelation function of the input signal. The advantage of this estimator is that it looks at the contribution of all the formant peaks within the autocorrelation function, thereby insuring an accurate pitch measurement of the stronger speech signal.

An example of the MLE method is provided in Figure 3.18. The upper left-hand plot is that of the original overlapping vocalic speech signal. The upper right-hand plot is the lowpass filtered signal and the bottom plot is the output of the periodic estimator. The pitch value is chosen as the maximum value of $g(P)$. The pitch estimate of the co-channel speech signal was 178 Hz which is very close to the measured pitch frequency, 176 Hz, of the stronger signal.

From our analysis, of the three techniques presented we have found the maximum likelihood pitch estimator to be least prone to error introduced from an interfering speech signal. In the following section we implement the MLE method into an iterative joint pitch estimation technique to measure the pitch contours of two speakers in a co-channel speaker environment.

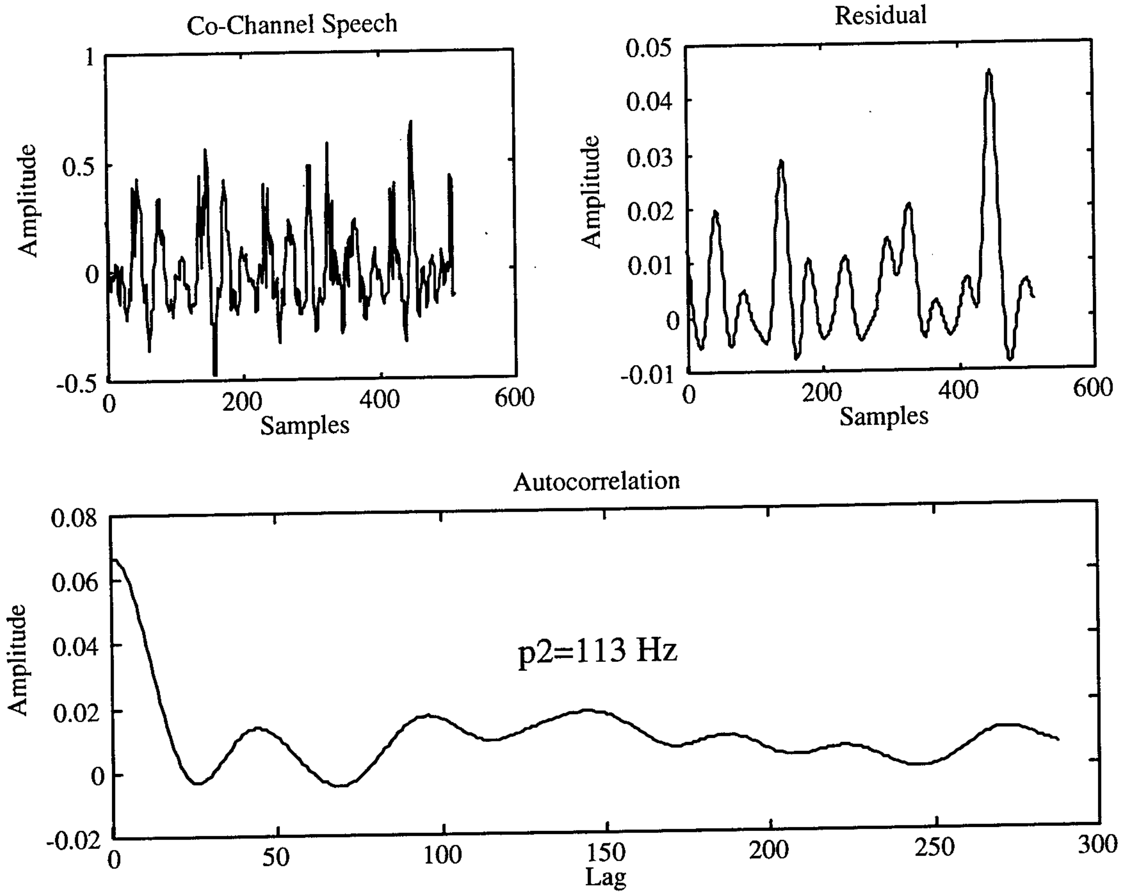


Figure 3.16: Modified autocorrelation pitch estimation on overlapping vocalic speech at SIR = -6 dB. Pitch estimated at 113 Hz.

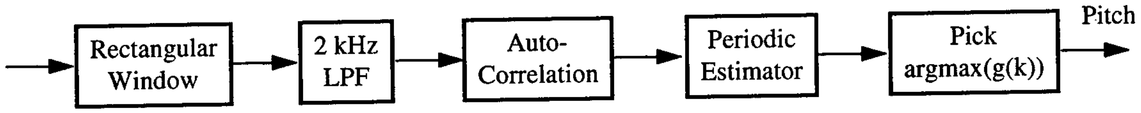


Figure 3.17: Maximum likelihood pitch estimation technique.

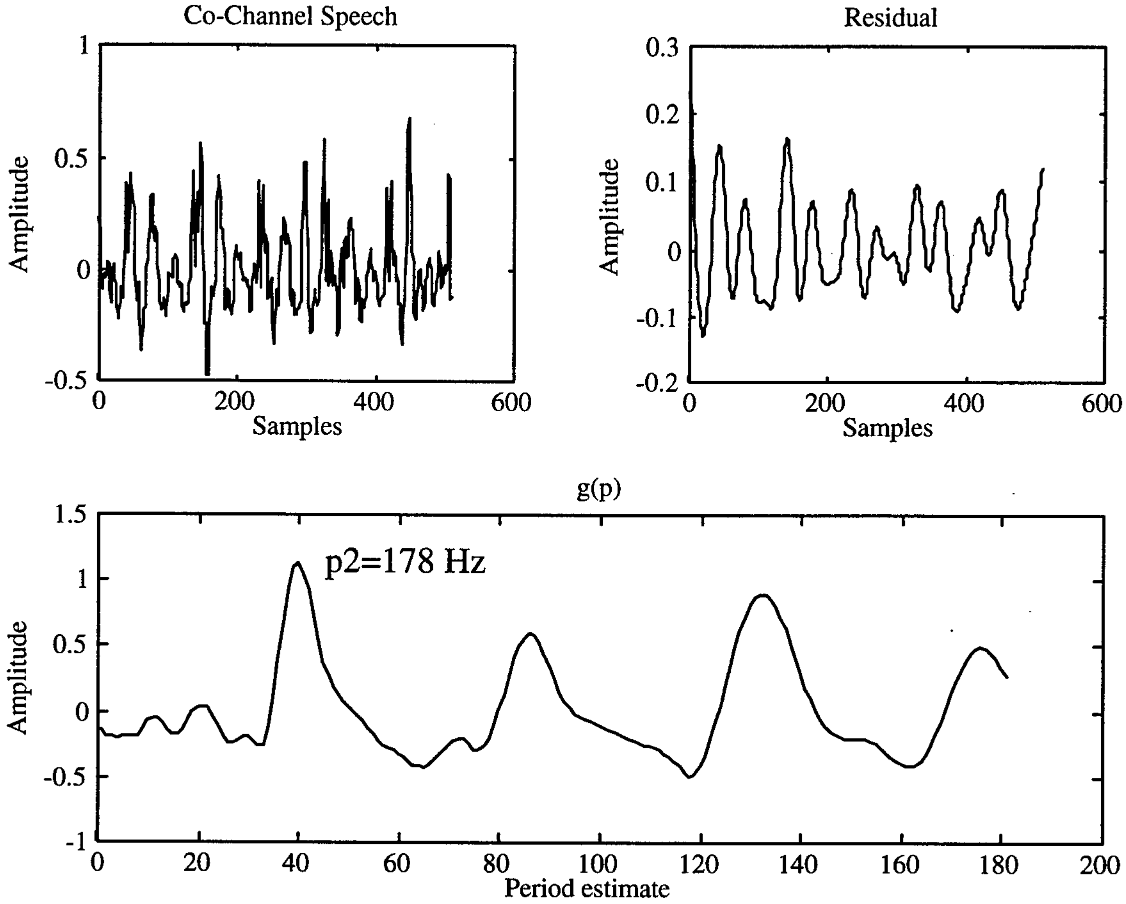


Figure 3.18: Maximum likelihood pitch estimation of overlapping vocalic speech at SIR = -6 dB. Pitch estimated at 178 Hz.

3.5.2 Maximum Likelihood Pitch Estimation with Harmonic Suppression

We have developed an iterative method to measure the pitch frequencies of co-channel speech using the MLE method and harmonic magnitude suppression. This technique first measures the F_0 formant of the stronger speaker using the maximum likelihood pitch estimator. Using this estimate, it then suppresses the harmonics associated with stronger speech signal using harmonic magnitude suppression. Our method of estimating the spectral harmonics of the stronger signal is provided in the next section. The MLE method is used on the residual signal to estimate the pitch of the weaker speech signal. This technique is outlined in Figure 3.19.

Referring to Figure 3.19, an estimate of the pitch of the stronger speech signal is measured using the MLE method. If there is another voiced speech signal present, this pitch estimate is used to suppress the spectral harmonics associated with the stronger signal. The MLE method is then applied to the residual signal to estimate the pitch of the weaker signal. These estimates are then compared to a average pitch values accumulated for each speaker. An assignment is made based on the minimum distance between the average pitch value for that speaker and the measured pitch estimate.

Figure 3.20 and Figure 3.21 shows an example resulting from applying our joint pitch estimation method to a voiced/voiced co-channel speech segment with a SIR = -6 dB. The pitch frequency of the stronger signal is correctly identified at 178 Hz. See Figure 3.20. The dominant harmonics associated with the stronger signal are suppressed and the signal is lowpass filtered, shown in the upper right-hand plot of Figure 3.21. The pitch frequency of the weaker signal is measured correctly at 111 Hz.

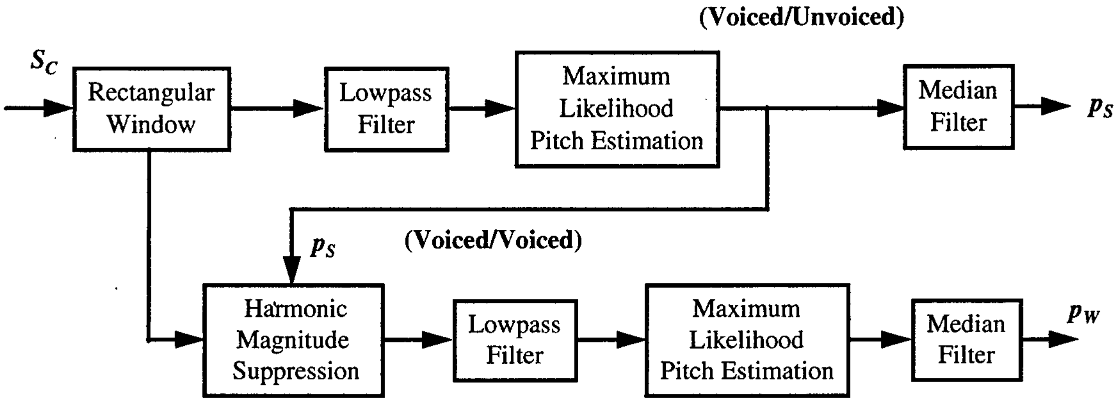


Figure 3.19: Joint pitch estimation using a maximum likelihood pitch estimator with harmonic magnitude suppression.

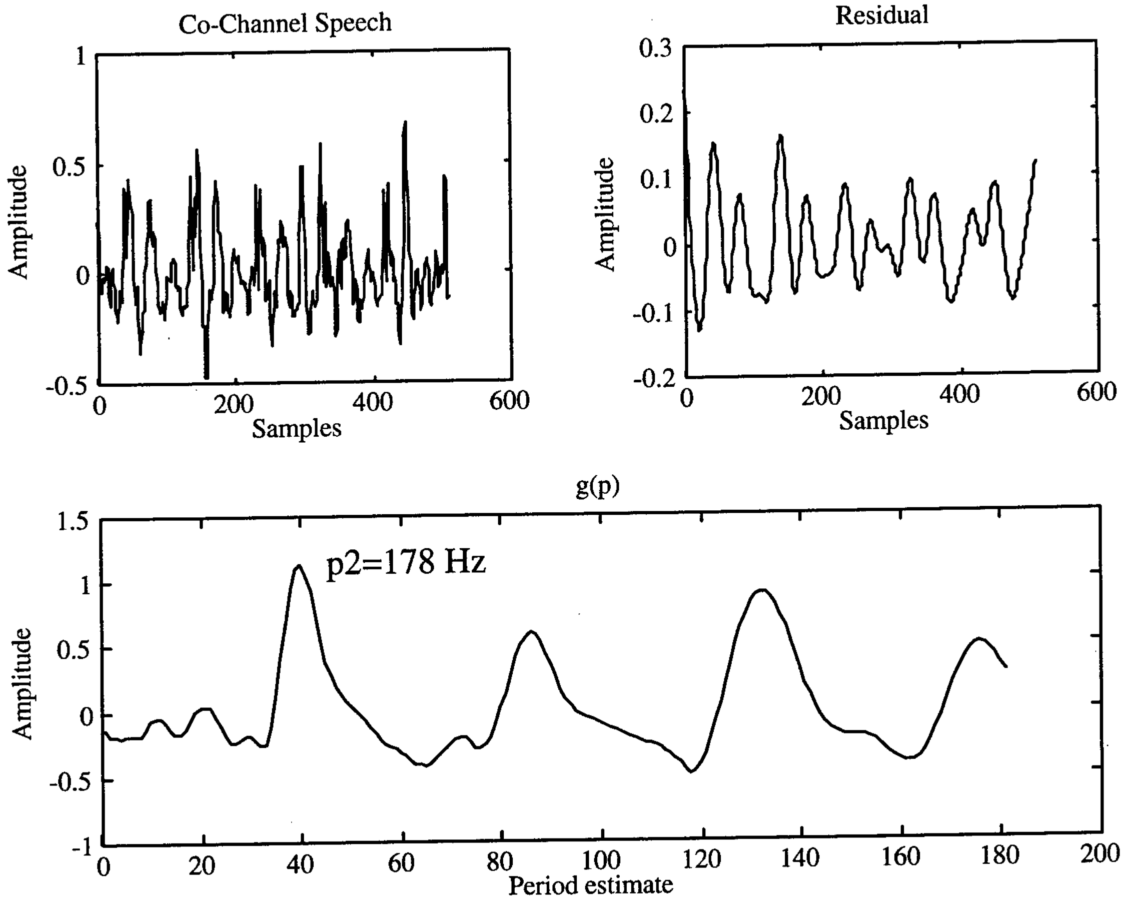


Figure 3.20: Maximum likelihood pitch estimation of overlapping vocalic speech at SIR = -6 dB. Pitch estimate of stronger signal measured at 178 Hz.

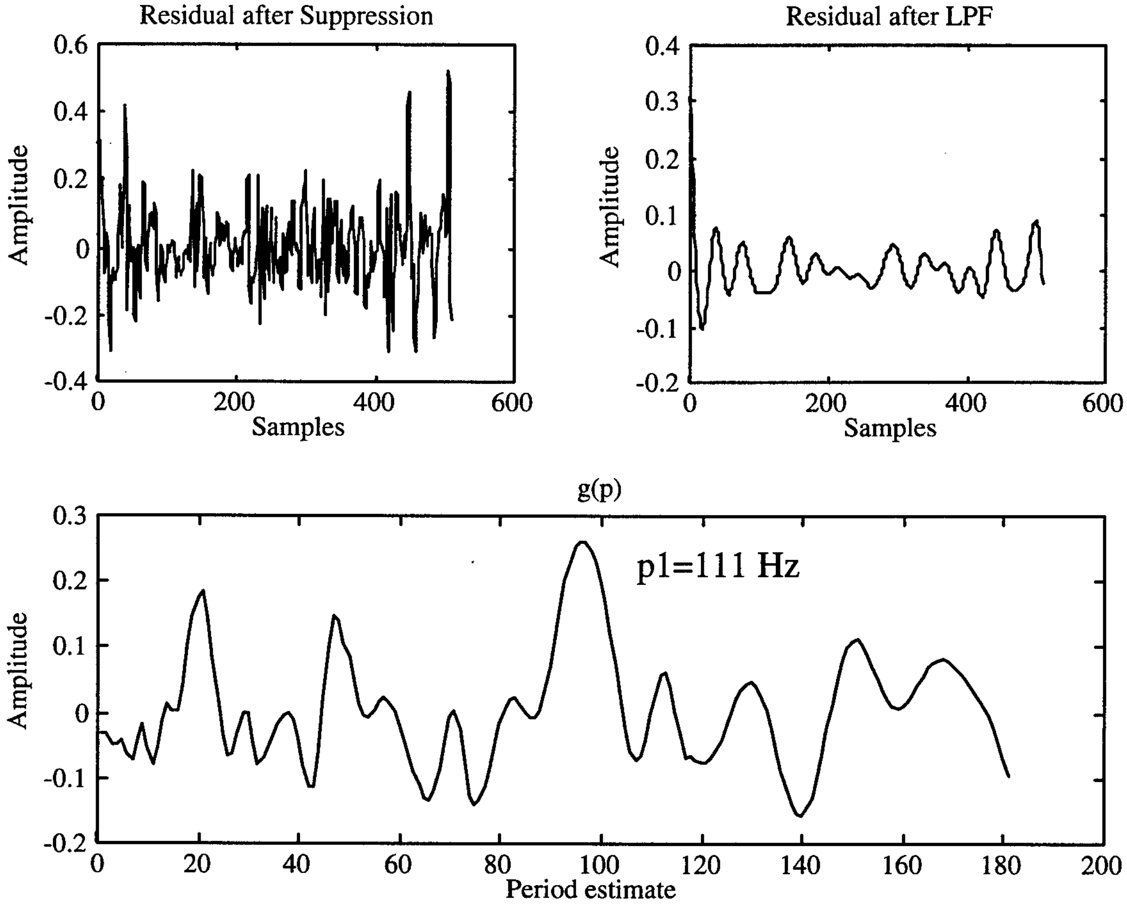


Figure 3.21: Maximum likelihood pitch estimation on residual co-channel speech after harmonic suppression of dominant harmonics. Pitch estimate of weaker signal is measured at 111 Hz.

3.6 Harmonic Selection

Given an ideal voiced speech segment, if the fundamental frequency is at F_0 then inspection of the magnitude spectrum should reveal harmonics located at F_0, $2F_0$, $3F_0$,.... However, a typical segment of voiced speech is only quasi-stationary. The center frequencies of these harmonics may not necessarily be at multiples of F_0. Also, there is further corruption of the center frequency positions due to the addition of multiple signals within the channel.

Predicting the center frequencies of harmonics using the pitch estimate has been widely used in the literature for harmonic suppression. A variation we have developed to this technique is as follows. First identify all peaks below a specified cut-off frequency within the co-channel speech magnitude spectrum. Using an estimate of the fundamental frequency of the stronger signal (F_{s0}), assign to the stronger signal those significant peaks with center frequencies that are at or near multiples of F_{s0} (F_{s0}, $2F_{s0}$, $3F_{s0}$,...). It is assumed the harmonics associated with the stronger signal will have the highest overall energy. Using an estimate of the fundamental frequency of the weaker signal (F_{w0}), assign to the weaker signal the remaining peaks, with center frequencies that are at or near multiples of F_{w0} (F_{w0}, $2F_{w0}$, $3F_{w0}$,...). Problems arise when a harmonic from the weaker signal overlaps a harmonic from the stronger signal. Under these conditions, we split the energy in that spectral peak between the two signals, giving the full energy of the spectral peak to the stronger signal and half the energy of that spectral peak to the weaker signal. If this value provides a smooth spectral magnitude contour for the weaker speech segment, the harmonic parameters for both speakers are used to initialize the optimization routine. If the contour is not smooth, then we switch values and assign the full energy of the spectral peak to the weaker signal and half the energy of the spectral peak to the stronger signal. This provides a robust method to effectively initialize the harmonic parameters for overlapping vocalic speech.

3.7 Speech Synthesis and Reconstruction

Speech can be thought of as a series of silence, voiced and unvoiced sounds concatenated together to form intelligible speech. In this research, voiced speech segments have been resynthesized using a sinusoidal representation. This model reconstructs the voiced sounds using measured parameters from the co-channel signal. The model relies on estimates of the amplitude, center frequency and phase of each harmonic associated with each voiced speech segment to produce natural and intelligible speech. The unvoiced speech segments are obtained directly from filtering the co-channel speech signal. The discrete co-channel speech waveform is segmented by sliding a Hanning weighted window over the co-channel speech with a 50% overlap. Final reconstruction to produce intelligible, natural sounding speech is performed by sequentially overlapping (by 50%) and adding the resynthesized speech segments for each particular speaker.

The ownership of a segment of speech, which has been separated from the co-channel speech segment is crucial to speech reconstruction. Misrepresentation of speech segments can adversely affect the intelligibility of the reconstructed speech signal. The use of pitch contours has been a widely accepted method of identifying ownership of voiced speech segments. We rely on the voicing state determination algorithm, discussed in the previous section, along with the pitch contour to assign ownership of the reconstructed speech segments.

We have shown in Chapter 2, that a Hanning weighted window is the best choice for segmenting the input data because of its preferred tradeoff of bandwidth versus leakage suppression. This tapered window is also compatible with the overlap-and-add processing used to reconstruct the signal at the output. In our case, we process the speech segments in the frequency domain, followed by an inverse transformation to the time domain. The inverse transform is modulated by the time-weighting window function. By processing overlapping frames, the tapered sections of the speech segments are added to produce natural sounding speech. This method smoothes any discontinuities resulting from the differences in processing consecutive frames.

4. RESULTS AND ANALYSIS

The algorithms developed in Chapter 3 have been tested independently to measure their performance. These algorithms have also been implemented and tested in a speaker separation system using co-channel speech data created from the TIMIT database and using co-channel speech recorded in our laboratory. In this chapter, we present results on the voicing determination algorithm, the joint pitch estimation algorithm, the nonlinear constrained least squared optimization algorithm and our simultaneous adaptive co-channel speaker separation system using both synthetic and real speech data.

4.1 Speech Databases

There currently is no standard database for co-channel speech signals. Such a database would have to consist of clean speech signals from a large set of speakers, along with a set of co-channel speech from these same speakers. The speech signals would have to be recorded on multiple channels, simultaneously. Therefore, we have created two co-channel speech databases. The first one uses speech signals taken from an existing database to produce co-channel speech. We have chosen speech from the TIMIT database. Our co-channel version of this database will be referred to as the TIMIT_COC database. The second speech database was created from our own recordings of speech signals from a male and female talker, both separately and in a co-channel environment. This database is labeled the RPI_COC database.

The TIMIT database consists of ten read English sentences from each of 640 different speakers along with a hand-segmented phone transcription of each utterance. The sampling rate is 16 kHz. The TIMIT database is considered a standard database and is used extensively in all forms of speech processing research. The speakers and sentences we have chosen to use in our analysis are provided in Table 4.1. Using the TIMIT database, co-channel speech signals were created by first normalizing each speech waveform based on the average energy per frame. Two speech signals were then summed on a computer, in the time domain, at a given SIR to produce the desired co-channel speech signal. We have created male/female, female/female and male/male speech mixtures at SIRs of 0, -6 and -12 dB. Table 4.2 identifies the speech sentences used for training and those used for testing. Table 4.3 - 4.5 shows the four co-channel speech sentence combinations used for testing. These speech mixtures closely resemble a co-channel speaker environment.

The RPI_COC database was recorded using an omni-directional dynamic microphone. The microphone has a frequency response between 100 and 8,000 Hz with a sensitivity of ±4 dB at 1,000 Hz. The data was recorded at a sampling rate of 22.050 kHz and digitized at 16 bits per sample. The data was then downsampled to 16 kHz, consistent with the TIMIT speech data. The speech consisted of English-native speaking male and female talker, reading from a manuscript, for four minutes. Speech data was recorded separately and simultaneously using a single microphone to create true co-channel speech signals. In the co-channel environment, two talkers were place equi-distance from the microphone and spoke simultaneously. This data was used to test the overall effectiveness of our speaker separation system in a realistic co-channel speaker environment.

Using the speech databases described above has allowed us to evaluate the performance of our VSDA and joint pitch estimation algorithms by comparing our results to reference signals obtained from the uncorrupted speech. It has also allowed us to compare the quality of our reconstructed speech signal to the original speech signal. Speech from the TIMIT database was used to perform parameter estimation analysis, including voicing state determination, joint pitch estimation and speaker separation.

Table 4.1: TIMIT database speech sentences used in analysis.

TIMIT Database (Northern USA) Speech Sentences			
Male		**Female**	
mbjv0	**mcew0**	**fajw0**	**Fcmm0**
sa1	sa1	sa1	sa1
sa2	sa2	sa2	sa2
si1247	si1442	si1263	si1083
si1877	si2072	si1893	si1957
si617	si812	si633	si453
sx167	sx182	sx183	sx183
sx257	sx272	sx273	sx273
sx347	sx362	sx3	sx363
sx437	sx452	sx363	sx420
sx77	sx92	sx93	sx93

Table 4.2: Speech sentences used to train and test the VSDA

Male		Female	
mbjv0	**mcew0**	**fajw0**	**Fcmm0**
Training Data			
sa2	sa2	sa1	sa1
sx167	sx182	sx183	sx183
sx257	sx272	sx273	sx273
sx347	sx362	sx3	sx363
sx437	sx452	sx363	sx420
sx77	sx92	sx93	sx93
Testing Data			
sa1	sa1	sa2	sa2
si1247	si1442	si1263	si1083
si1877	si2072	si1893	si1957
si617	si812	si633	si453

Table 4.3: Speech sentence pairs used for testing male/female co-channel speech.

Male/Female		
Co-Channel Sentence	**Talker 1 mbjv0**	**Talker 2 fajw0**
C1	sa1	sa2
C2	si1247	si1893
C3	si1877	si1263
C4	si617	si633

Table 4.4: Speech sentence pairs used for testing male/male co-channel speech.

Male/Male		
Co-Channel **Sentence**	**Talker 1** **mbjv0**	**Talker 2** **mcew0**
C1	sa1	si812
C2	si1247	si2072
C3	si1877	si1442
C4	si617	sa1

Table 4.5: Speech sentence pairs used for testing female/female co-channel speech.

Female/Female		
Co-Channel **Sentence**	**Talker 2** **fajw0**	**Talker 2** **fcmm0**
C1	sa2	sa2
C2	si1263	si1957
C3	si1893	si1083
C4	si633	si453

4.2 Parameter Estimation Analysis

This section presents results obtained from testing the voicing state determination algorithm (VSDA) and the joint pitch estimation algorithm on the TIMIT_COC speech database.

4.2.1 Voicing State Determination Algorithm

Three different sets of co-channel speech mixtures were used to test our co-channel VSDA. The spoken language was English and the signals were taken from the TIMIT_COC database. The signals were extracted from the database and combined in a male/female, male/male and female/female mixture as described above. All three mixtures were tested at SIR = 0 dB, with the male/female mixture also tested at SIR = -6 dB.

From the database, we chose six sentences for training and four sentences for testing. See Table 4.2. The training sentences were specifically chosen to provide representation of the different phones produced during speech. The test sentences provided a reasonable sampling of the different phones that would be found in normal speech. The true voicing state for each signal was determined prior to mixing.

Voicing state determination of an uncorrupted speech signal is based on the zero crossing rate and the short time energy measure. The convention we adopted for classifying the voicing state of uncorrupted speech is as follows. When the zero crossing rate was equal to or below the threshold and the energy measure was above the threshold, the speech was label as *voiced*. If the zero crossing rate was above the threshold and the energy measure was also above the threshold, the speech segment was labeled as *unvoiced*. Otherwise, the energy was below the threshold and the speech segment was labeled as *silence*. This was the same technique we used to create the training data for the VSDA. The *a priori* probability of a particular state occurring was set to (Refer to Figure 3.12):

Level 1:	$P(\textit{some voiced speech}) = .85$	$P(\textit{unvoiced speech}) = .15$
Level 2:	$P(\textit{all voiced speech}) = .5$	$P(\textit{mixed voiced speech}) = .5$
Level 3:	$P(\textit{voiced/unvoiced speech}) = .5$	$P(\textit{unvoiced/voiced speech}) = .5$

assuming the two speakers were speaking simultaneously for the total length of time.

The training data for each mixture was formed by segmenting the six test sentences into voiced, unvoiced, and silent frames. The voiced, unvoiced, and silence frames were randomly mixed with the other speakers voiced, unvoiced, and silence frames to produce 2000 voiced/voiced, voiced/unvoiced, unvoiced/voiced, and unvoiced/unvoiced co-channel speech frames each. The co-channel speech segments were then grouped according to the classifications outlined in Figure 3.12. Feature vectors were extracted from the training data. The Fisher linear discriminant was applied to the feature set to project them into the single dimension which provided the best separation. This was performed at each of the three classification levels described above.

The Bayes, *kNN*, and Parzen window classifiers, outlined in Section 3.4.4 through Section 3.4.6, where applied to the sample data to determine the voicing classification.

In the Bayes approach, the thresholds were calculated using the Bayes criterion, assuming a normal distribution function. In the *kNN* approach, the number of nearest neighbors was set to $k = 7$. We used a uniform kernel function for the Parzen window and a volume or window width approximately equal to the standard deviation of the sampled data. The volume or window width, in the Parzen window approach, varied at each decision level. As with the number of nearest neighbors, several experiments were run to determine the optimal window size. While all the distributions were not normal, all were unimodal. In the Bayes approach using the assumption that the distributions were normal, provided similar results to the nonparametric approaches. Histograms of the sample data used in the analysis, for a male/female speech mixtures at 0 dB are provided in Figure 4.1 through Figure 4.3.

Cumulative results of all four test sentence combinations for each speech mixture are presented below. A confusion matrix of the results from the voicing state determination algorithm, for the male/female, male/male and female/female speech mixtures at SIR = 0 dB are presented in Tables 4.6 - 4.14 using the three approaches given above. The results for the male/female mixture at SIR = -6 dB are presented in Table 4.15 using only the Bayes approach. The results are in percent detection, with the raw scores (number of frames detected) provided in parentheses. The values along the main diagonal in the tables represent the percentage of co-channel speech segments labeled correctly. The values down each column, off the main diagonal, represent the percentage of incorrectly labeled segments belonging to the true state (missed detection).

The values along each row, off the main diagonal, represent the percentage of speech segments which were incorrectly identified belonging to the estimated state (false detect).

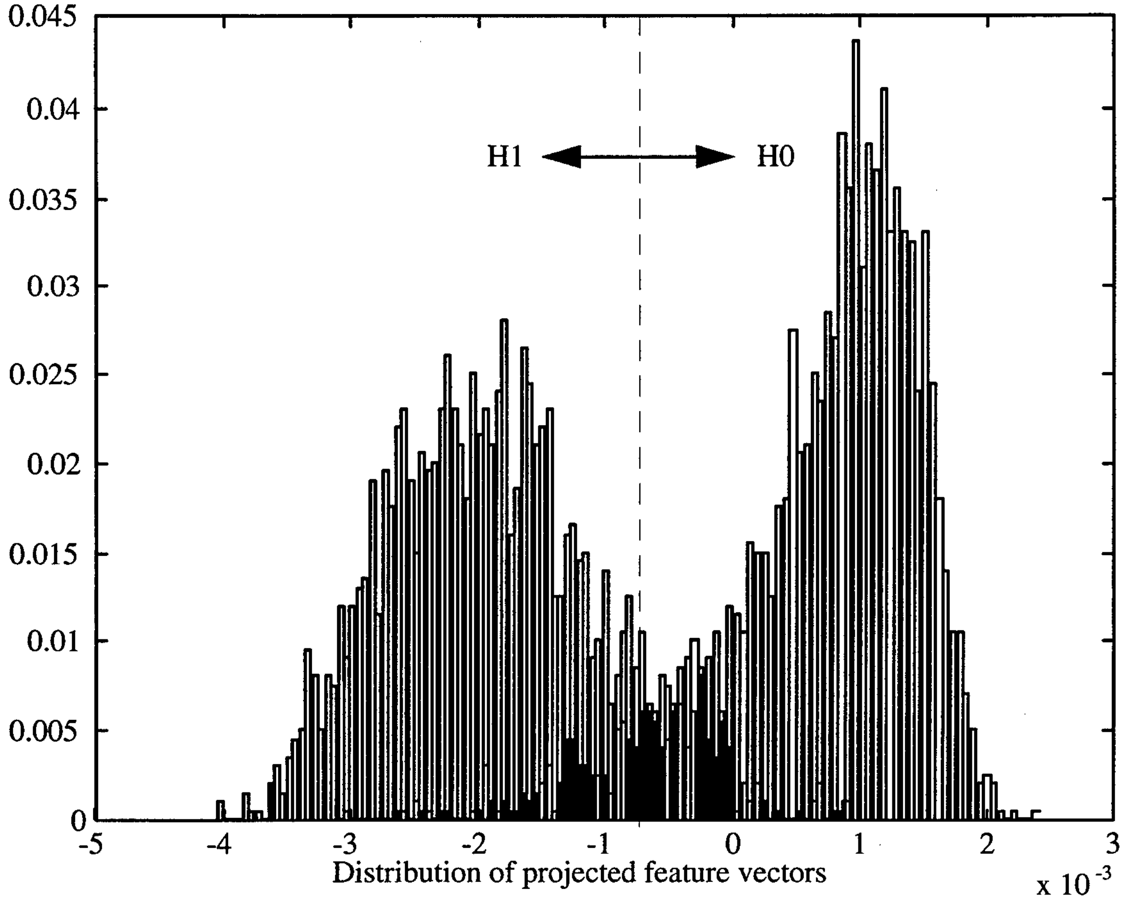


Figure 4.1: Histogram comparing distribution between Voiced (H0) and Unvoiced (H1) feature values used in training (Level 1) on a male/female speech mixture.

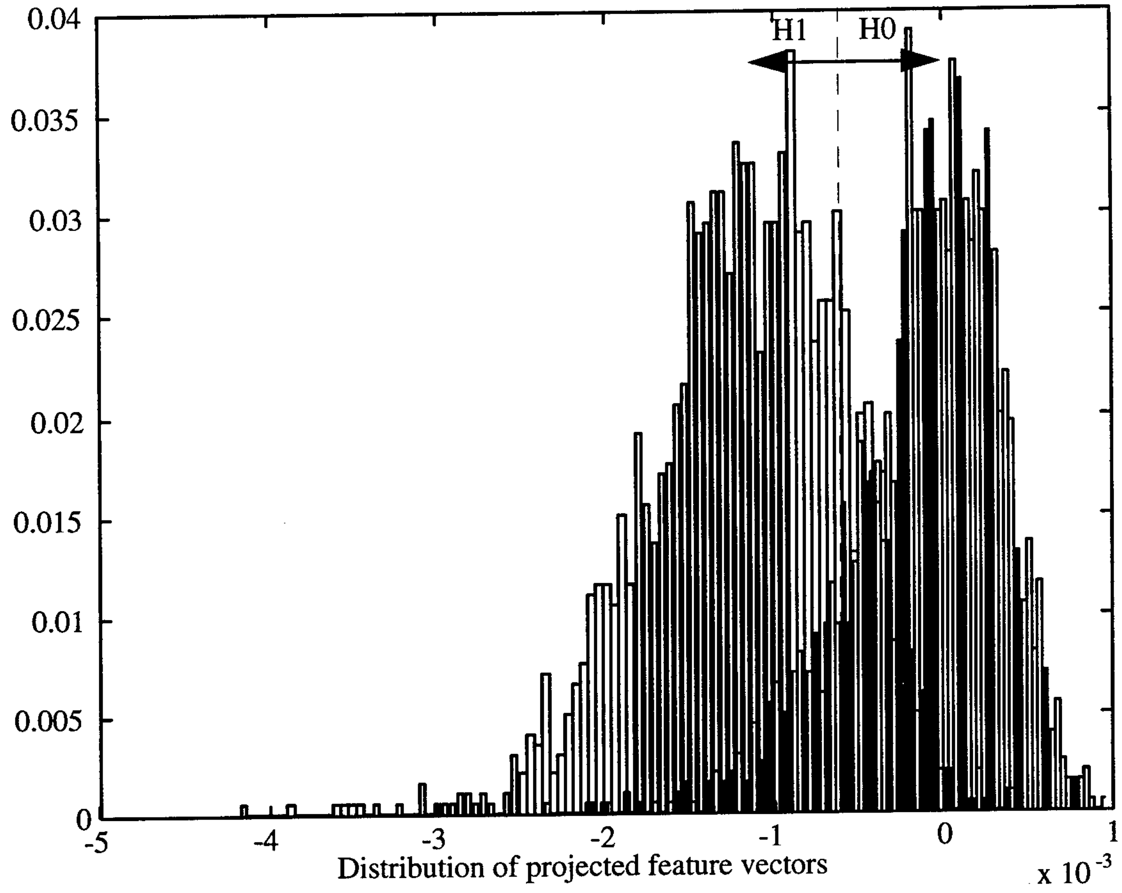


Figure 4.2: Histogram displaying distribution of All Voiced (H0) and Mixed Voiced (H1) feature values used in training (Level 2) on a male/female speech mixture.

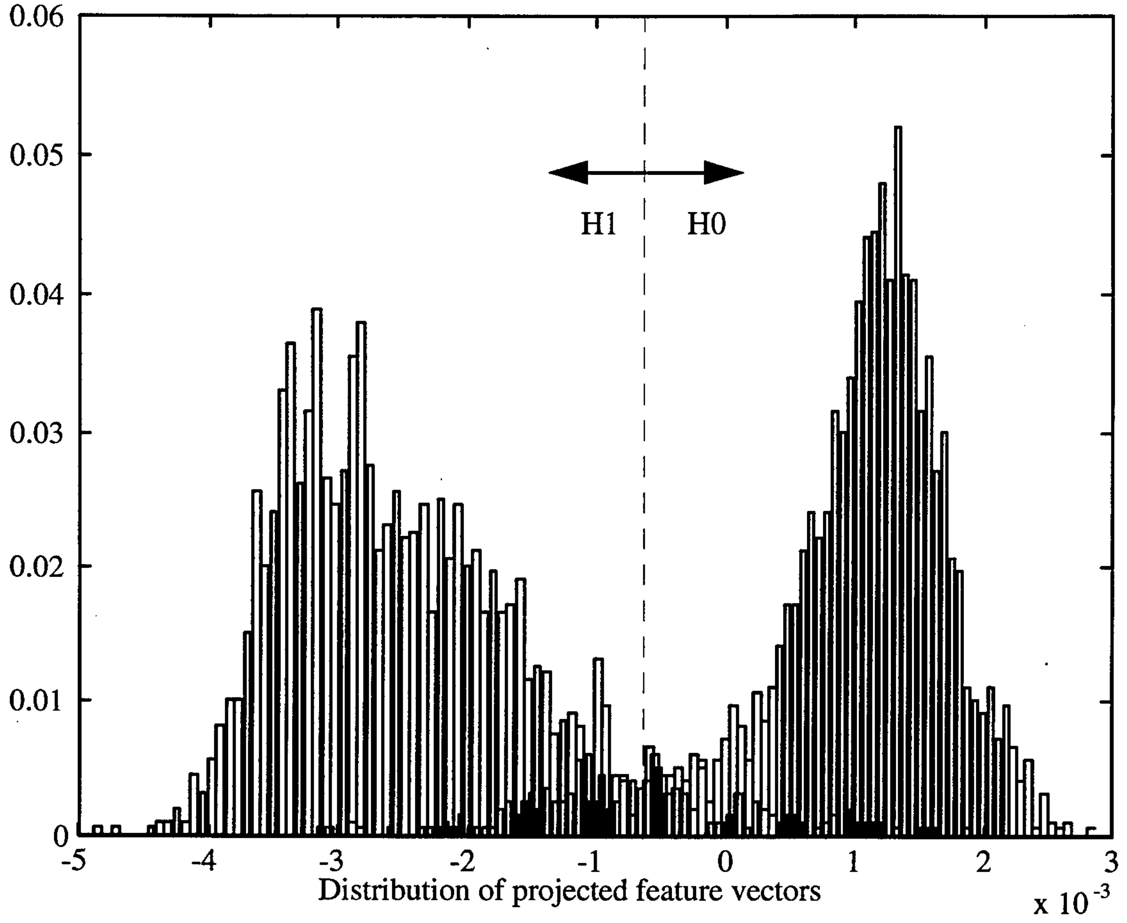


Figure 4.3: Histogram displaying distribution of Voiced/Unvoiced (H0) and Unvoiced/Voiced (H1) feature values used in training (Level 3) on a male/female speech mixture.

The VSDA performed better on the different gender mixtures than on the same gender mixtures. Speech characteristics vary more between speakers of different gender than speakers of the same gender. The three classifiers had similar results in overall detection rates. Considering just the results from the Bayes classifier, the overall classification performance of the male/female speech mixtures was 83.43%, while that for the male/male mixtures was 76.49% and for the female/female mixtures was 71.99%. For the different gender mixtures, the performance of the VSDA improved to 84.21% when the SIR was -6 dB compared to the same speech sentence pairs at SIR = 0 dB. This slight improvement of the overall detection performance was mainly due to the increased classification of the stronger speaker's voicing states.

The majority of the errors, consistent throughout each speech mixture and each approach, occurred when mixed voiced speech (V/UV and UV/V) was misclassified as all voiced speech (V/V). Frames in which there is a voicing transition (onset or offset of voicing) resulted in mixed voiced speech being misclassified as all voiced or all unvoiced.

Table 4.6: Confusion matrix, using the Bayes classifier, of Male/Female speech mixtures with SIR = 0 dB. Overall 83.43% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.

Voicing State	**SIL**	**V/V**	**V/UV**	**UV/V**	**UV/UV**
SIL	100 (86)	0	0	0.34 (1)	6.44 (15)
V/V	0	84.17 (234)	24.11 (34)	7.82 (23)	0
V/UV	0	10.43 (29)	63.12 (89)	1.02 (3)	2.58 (6)
UV/V	0	5.40 (15)	1.42 (2)	88.10 (259)	8.15 (19)
UV/UV	0	0	11.35 (16)	2.72 (8)	82.83 (193)

Table 4.7: Confusion matrix, using the *kNN* classifier, of Male/Female speech mixtures with SIR = 0 dB. Overall 83.62% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.

Voicing State	**SIL**	**V/V**	**V/UV**	**UV/V**	**UV/UV**
SIL	100(86)	0	0	0.34(1)	6.01(14)
V/V	0	79.50(221)	14.18(20)	5.44(16)	0
V/UV	0	13.31(37)	66.67(94)	1.02 (3)	1.72(4)
UV/V	0	7.19(20)	0	87.07(256)	3.86(9)
UV/UV	0	0	19.15(27)	6.12(18)	88.41(206)

Table 4.8: Confusion matrix, using the Parzen window classifier, of Male/Female speech mixtures with SIR = 0 dB. Overall 83.24% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.

Voicing State	SIL	V/V	V/UV	UV/V	UV/UV
SIL	100 (86)	0	0	0.34(1)	6.01(14)
V/V	0	80.22(223)	13.48(19)	4.76(14)	0.43(1)
V/UV	0	12.95(36)	80.84(114)	1.70(5)	5.58(13)
UV/V	0	6.83(19)	2.13(3)	91.84(270)	16.74(39)
UV/UV	0	0	3.55(5)	1.36(4)	71.24(166)

Table 4.9: Confusion matrix, using the Bayes classifier, of Male/Male speech mixtures with SIR = 0 dB. Overall 76.49% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.

Voicing State	SIL	V/V	V/UV	UV/V	UV/UV
SIL	100 (137)	0	0	0	10.08 (25)
V/V	0	84.39 (265)	23.80 (25)	10.66 (29)	1.21 (3)
V/UV	0	9.56 (30)	47.62 (50)	12.87 (35)	2.42 (6)
UV/V	0	6.05 (19)	10.48 (11)	66.18 (180)	9.27 (23)
UV/UV	0	0	18.10 (19)	10.29 (28)	77.02 (191)

Table 4.10: Confusion matrix, using the *kNN* classifier, of Male/Male speech mixtures with SIR = 0 dB. Overall 74.16% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.

Voicing State	**SIL**	**V/V**	**V/UV**	**UV/V**	**UV/UV**
SIL	100 (137)	0	0	0	10.08 (25)
V/V	0	78.34(246)	19.05(20)	9.56(26)	1.21 (3)
V/UV	0	12.42(39)	38.10(40)	9.93(27)	1.61(4)
UV/V	0	8.60(27)	8.57(9)	66.91(182)	9.27 (23)
UV/UV	0	0.64(2)	34.28(36)	13.60(37)	77.83(193)

Table 4.11: Confusion matrix, using the Parzen window classifier, of Male/Male speech mixtures with SIR = 0 dB. Overall 77.42% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.

Voicing State	**SIL**	**V/V**	**V/UV**	**UV/V**	**UV/UV**
SIL	100 (137)	0	0	0	10.08 (25)
V/V	0	79.94(251)	14.28(15)	5.88(16)	1.21(3)
V/UV	0	11.46(36)	62.86(66)	11.40(31)	4.03(10)
UV/V	0	8.60(27)	11.43(12)	77.94(212)	16.94(42)
UV/UV	0	0	11.43(12)	4.78(13)	67.74(168)

Table 4.12: Confusion matrix, using the Bayes classifier, of Female/Female speech mixtures with SIR = 0 dB. Overall 71.99% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.

Voicing State	**SIL**	**V/V**	**V/UV**	**UV/V**	**UV/UV**
SIL	98.75 (79)	0	0	1.24 (2)	8.57 (18)
V/V	0	81.51 (260)	23.90 (60)	13.66 (22)	0
V/UV	0	7.21 (23)	54.59 (137)	24.84 (40)	4.76 (10)
UV/V	0	10.97 (35)	13.94 (35)	50.94 (82)	2.38 (5)
UV/UV	1.25 (1)	0.31 (1)	7.57 (19)	9.32 (15)	84.29 (177)

Table 4.13: Confusion matrix, using the *kNN* classifier, of Female/Female speech mixtures with SIR = 0 dB. Overall 71.79% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.

Voicing State	**SIL**	**V/V**	**V/UV**	**UV/V**	**UV/UV**
SIL	98.75 (79)	0	0	1.24 (2)	8.57 (18)
V/V	0	79.62(254)	21.12(53)	9.94(16)	0
V/UV	0	6.27(20)	55.78(140)	20.50(33)	3.33(7)
UV/V	0	13.48(43)	11.55(29)	47.83(77)	0.95(2)
UV/UV	1.25 (1)	0.63(2)	11.55(29)	20.50(33)	87.15(183)

Table 4.14: Confusion matrix, using the Parzen window classifier, of Female/Female speech mixtures with SIR = 0 dB. Overall 72.58% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.

Voicing State	**SIL**	**V/V**	**V/UV**	**UV/V**	**UV/UV**
SIL	98.75 (79)	0	0	1.24 (2)	8.57 (18)
V/V	0	77.43(247)	17.53(44)	9.32(15)	0
V/UV	0	9.09(29)	64.54(162)	23.60(38)	10.95(23)
UV/V	0	13.48(43)	15.54(39)	60.87(98)	6.67(14)
UV/UV	1.25 (1)	0	2.39(6)	4.97(8)	73.81(155)

Table 4.15: Confusion matrix, using the Bayes classifier, of Male/Female speech mixtures with SIR = -6 dB. Overall 84.21% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.

Voicing State	**SIL**	**V/V**	**V/UV**	**UV/V**	**UV/UV**
SIL	100 (75)	0	0	0	5.74 (14)
V/V	0	83.46 (232)	8.51 (12)	10.20 (30)	0
V/UV	0	10.07 (28)	74.47 (105)	1.02 (3)	2.87 (7)
UV/V	0	6.47 (18)	1.42 (2)	85.72 (252)	7.37 (18)
UV/UV	0	0	15.60 (22)	3.06 (9)	84.02 (205)

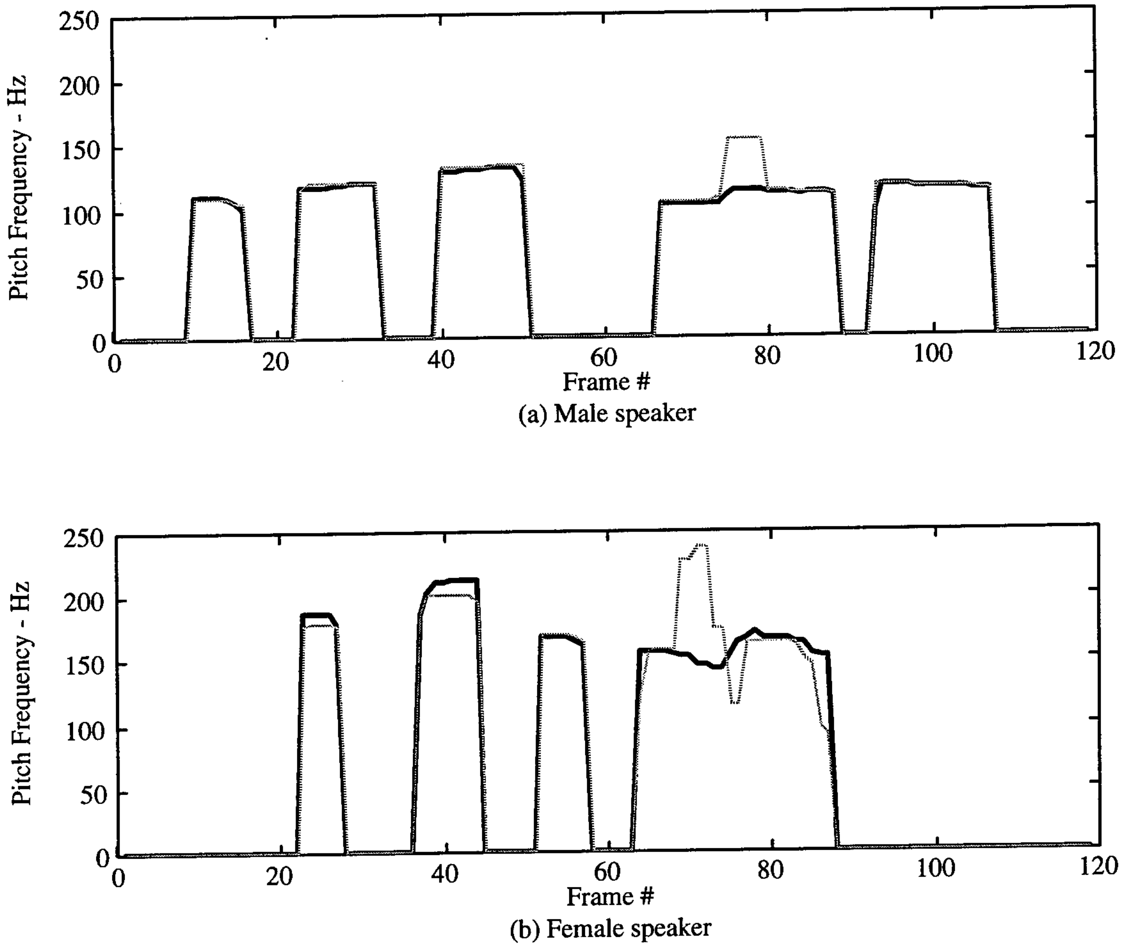
4.2.2 Joint Pitch Estimation

We have tested our joint pitch estimation algorithm, based on the maximum likelihood pitch estimator and harmonic suppression, on the same speech mixtures used to test the VSDA. For this test, we assumed the voicing state and gender of each speaker was known prior to pitch estimation. This information allowed us to isolate the performance of the pitch estimation routine.

The pitch contour of the co-channel speech was measured first, using the maximum likelihood pitch estimator, presented in Section 3.5.2. These values represented the pitch frequency of the stronger speech segment, which may or may not originate from the stronger speaker. The pitch contour was then passed through a median filter for smoothing. Based on the voicing state of the two speakers, the pitch of the weaker signal was calculated on those speech segments that were identified as voiced/voiced. The pitch frequency of the weaker signal was measured using harmonic suppression and the maximum likelihood pitch estimator. Refer back to Figure 3.19. For those segments, which were a mixture of voiced and unvoiced speech, the measured pitch frequency was assigned to the speaker which was voiced. The pitch contours were constructed on a frame-by-frame basis. The final pitch contours for both speakers were then passed through a median filter to remove any pitch doubling associated with the onset or offset of voicing. These values were then compared to the reference pitch contour, measured from the uncorrupted speech signal (prior to mixing) using the maximum likelihood pitch estimator. These pitch contours were passed through a median filter, again to reduce pitch doubling errors. An example comparing the reference pitch contour (measured from the uncorrupted speech signal) with the pitch contour measured from the co-channel speech signal is given in Figure 4.4.

We conducted experiments using the same four, two speaker co-channel speech sentences pairs used to test the VSDA. Cumulative results the four co-channel speech sentences for each speech mixture are presented below. Figures 4.5 - 4.8 provides a histogram of the pitch errors made by the joint pitch estimator accumulated over all four sentences for each of the four speech mixtures (M/F 0dB, M/M 0dB, F/F 0dB, and M/F -6dB). The pitch error was defined to be the difference between the reference pitch frequency and the estimated pitch frequency, relative to 100 Hz, for each frame. Tables 4.16 - 4.19 provides quantitative measurements that can be used to summarize these results by showing the percentage of frames in which the pitch error was greater than 20%, 10%, and 5% of our normalization factor (100 Hz). That is, the percentage of frames in which the pitch estimates were greater than 20 Hz, 10 Hz, and 5 Hz from the reference value. For these tests, we have evaluated the performance of the joint pitch estimator under realistic conditions by considering voiced/voiced, voiced/unvoiced, and unvoiced/voiced co-channel speech mixtures.

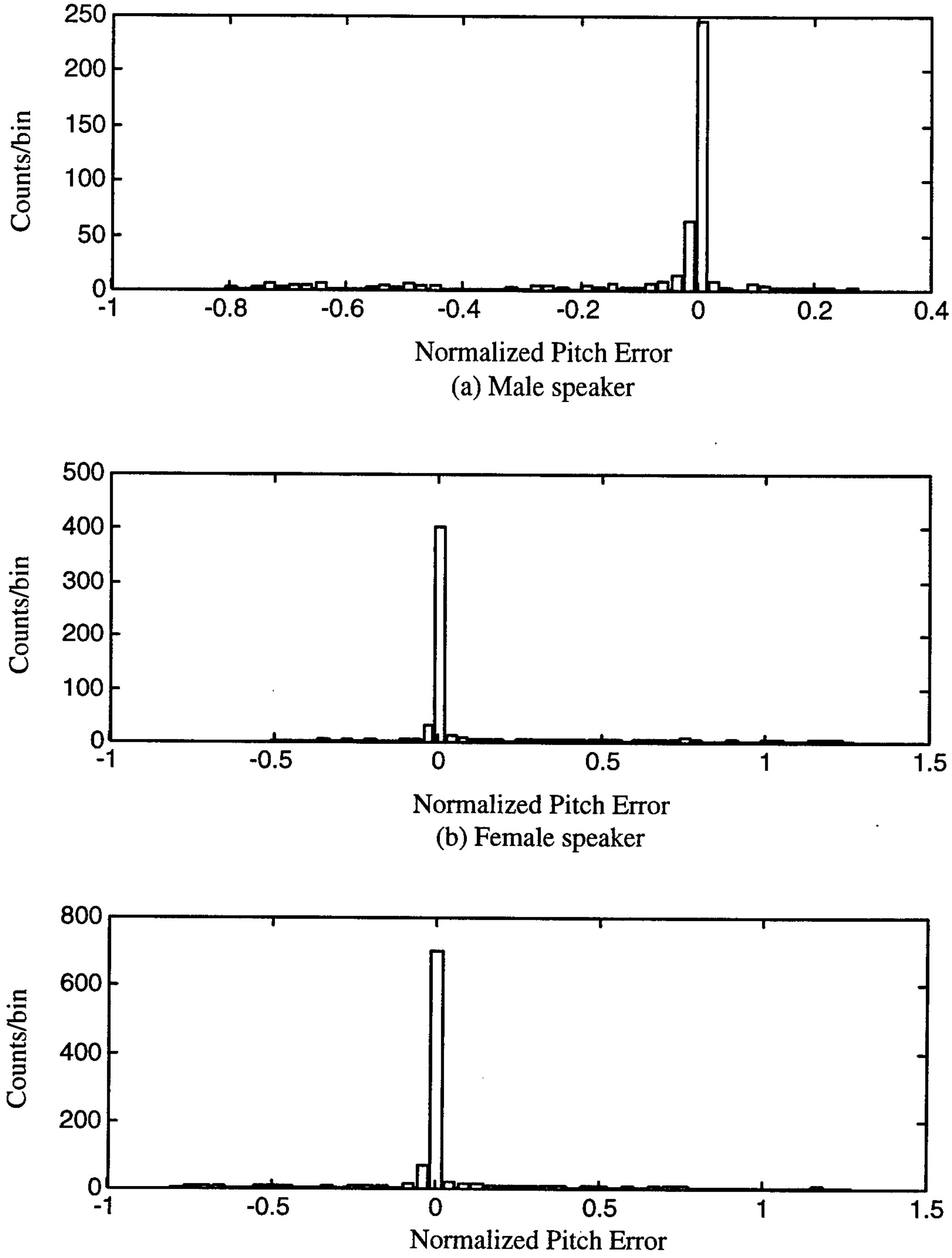
Pitch errors due to pitch doubling, occurred during the onset and offset of voicing. Pitch doubling occurs when the estimated pitch frequency is measured at twice the true pitch frequency. Pitch doubling contributed to those errors greater than 20 Hz. Pitch errors also resulted from pitch crossing, in which the pitch contour of one speaker crosses the pitch contour of another speaker. Poor suppression of the stronger harmonics also resulted in pitch errors by introducing erroneous harmonics in the residual signal that was used to estimate the pitch frequency of the weaker speaker. This caused pitch estimates of the weaker signal to be measured near the pitch estimates of the stronger signal. We can see that in the estimated pitch contour of the male speaker in Figure 4.4 (a).



(a) Male speaker

(b) Female speaker

Figure 4.4: Comparison between pitch contour measured on uncorrupted speech (solid line) and the pitch contour measured from the co-channel speech (dotted line) for (a) male speaker, and (b) female speaker. Speech was taken from the TIMIT_COC database and mixed at SIR = 0 dB.

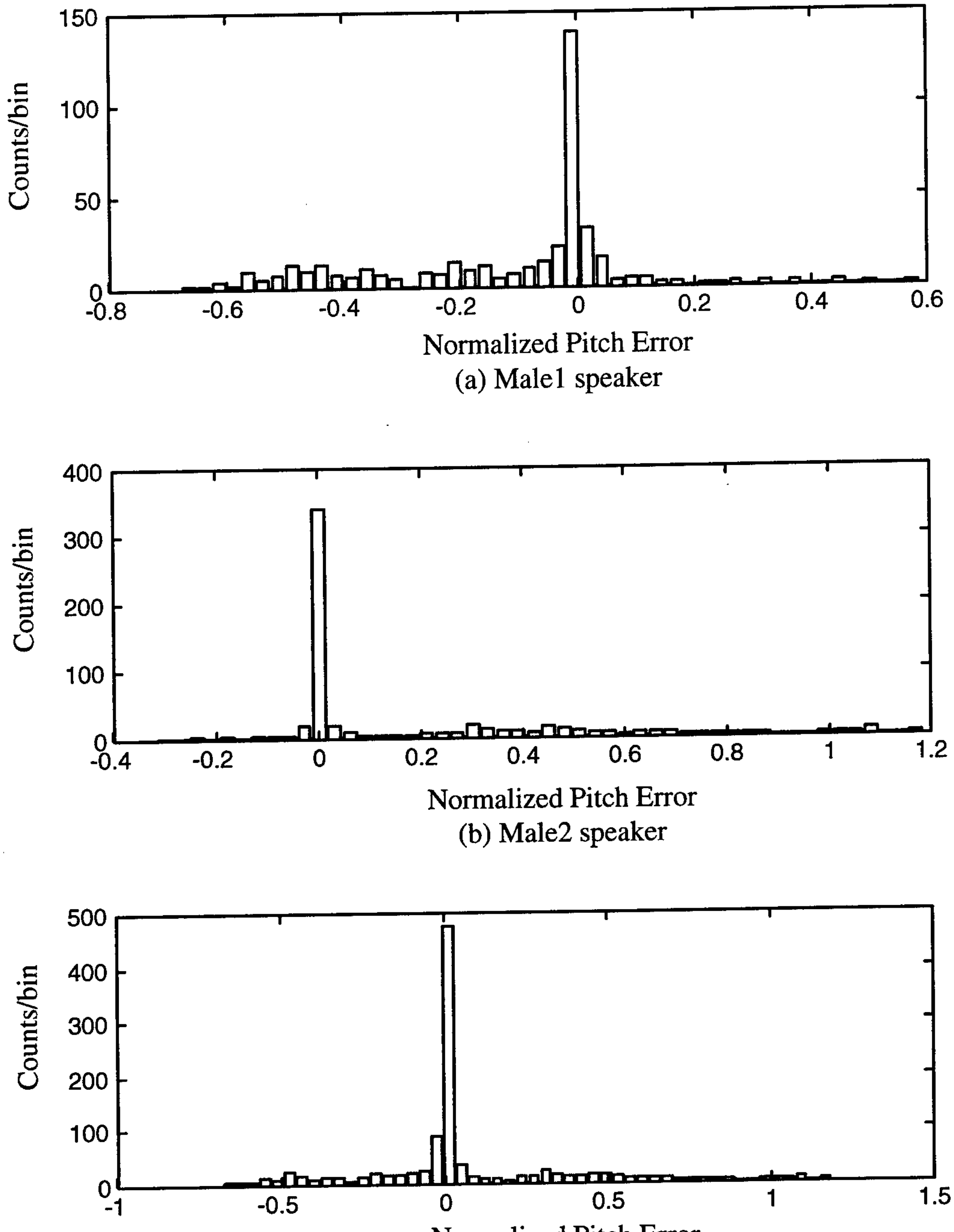


(a) Male speaker

(b) Female speaker

(c) Total for both speakers

Figure 4.5: Histogram of normalized pitch errors (relative to 100 Hz) at SIR = 0 dB for male/female co-channel speech mixtures. (a) Male, (b) Female, and (c) Total for both male and female pitch errors.

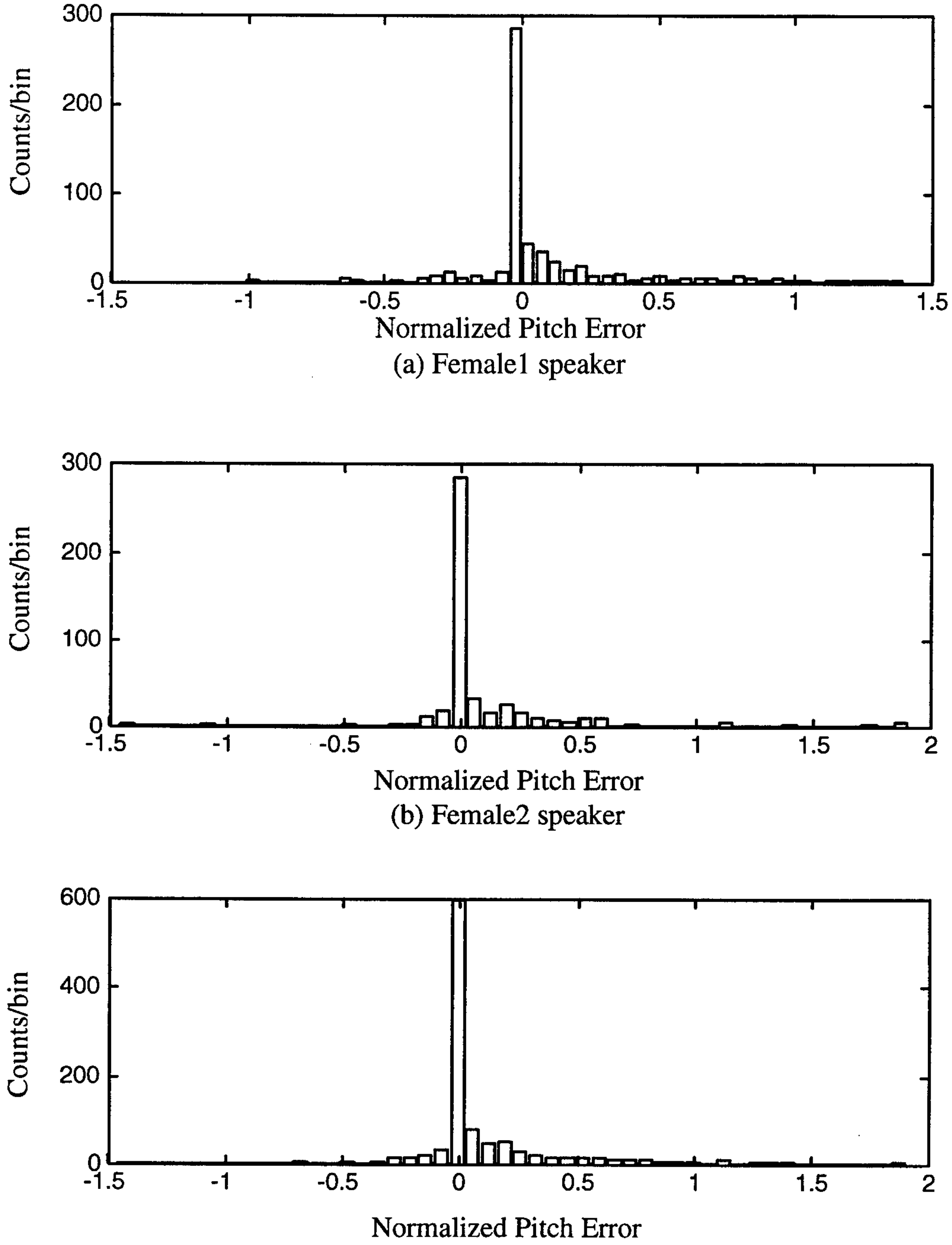


(a) Male1 speaker

(b) Male2 speaker

(c) Total for both speakers

Figure 4.6: Histogram of normalized pitch errors (relative to 100 Hz) at SIR = 0 dB for male/male co-channel speech mixtures. (a) Male1, (b) Male2, and (c) Total of both Male1 and Male2 pitch errors.

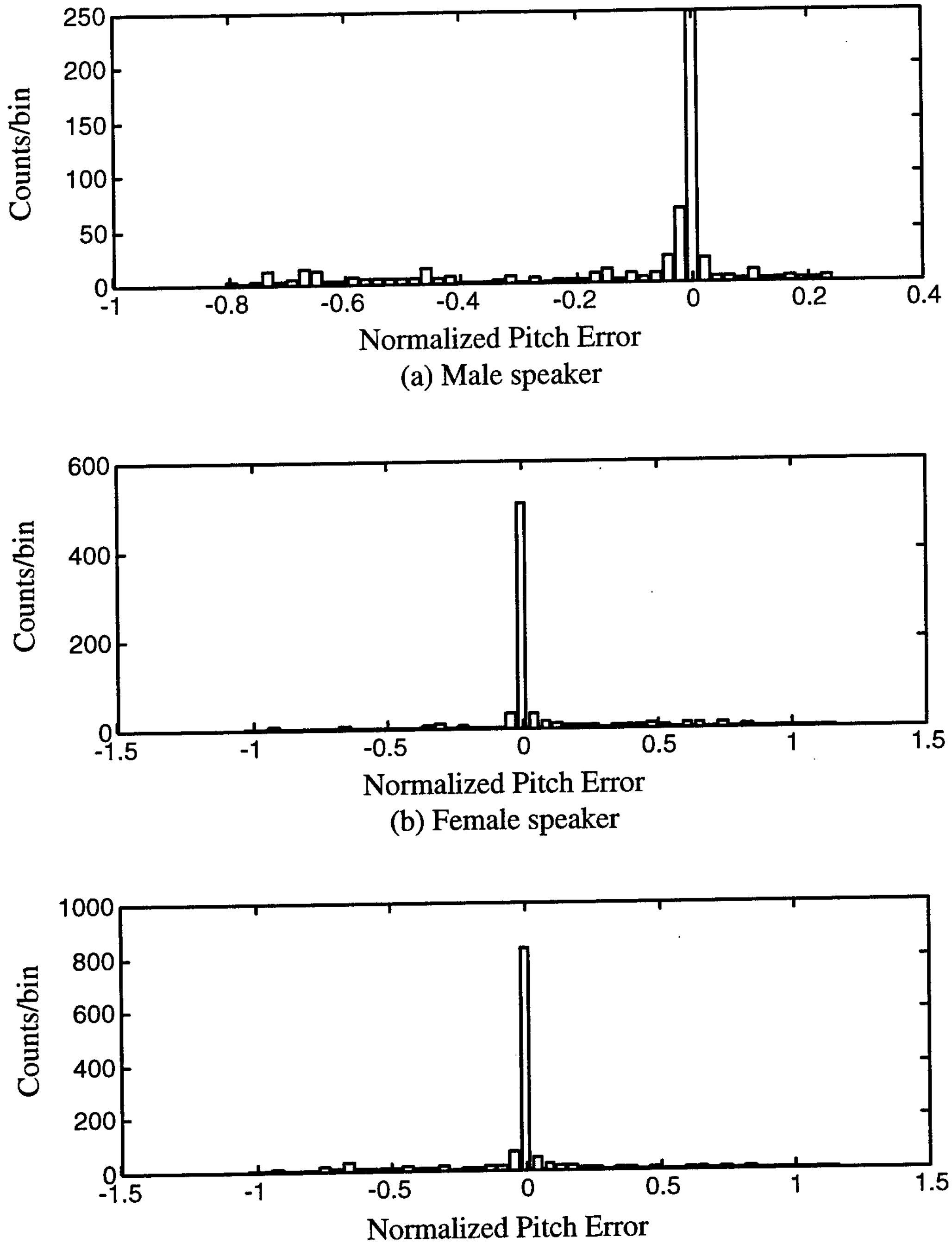


(a) Female1 speaker

(b) Female2 speaker

(c) Total for both speakers

Figure 4.7: Histogram of normalized pitch errors (relative to 100 Hz) at SIR = 0 dB for female/female co-channel speech mixtures. (a) Female1, (b) Female2, and (c) Total of both Female1 and Female2 pitch errors.



(a) Male speaker

(b) Female speaker

(c) Total for both speakers

Figure 4.8: Histogram of normalized pitch errors (relative to 100 Hz) at SIR = -6 dB for male/female co-channel speech mixture. (a) Male, (b) Female, and (c) Total of both male and female pitch errors.

Table 4.16: Percentage of pitch errors, for a male/female mixture (SIR = 0 dB), which were greater than 20%, 10%, and 5% of the normalization factor (100 Hz). The average pitch frequencies for the male and female speakers were 109.37 Hz and 186.44 Hz respectively.

Speaker	**Percentage of Pitch Errors greater than x% of 100 Hz**		
	>20%	**>10%**	**>5%**
Male	12.65	17.66	21.00
Female	17.69	19.79	22.42
Combined	15.56	18.89	21.82

Table 4.17: Percentage of pitch errors, for a male/male mixture (SIR = 0 dB), which were greater than 20%, 10%, and 5% of the normalization factor (100 Hz). The average pitch frequencies for the Male1 and Male2 speakers were 109.37 Hz and 157.0 Hz respectively.

Speaker	**Percentage of Pitch Errors greater than x% of 100 Hz**		
	>20%	**>10%**	**>5%**
Male1	31.5	42.24	48.69
Male2	30.38	33.11	36.01
Combined	30.85	36.92	41.29

Table 4.18: Percentage of pitch errors, for a female/female mixture (SIR = 0 dB), which were greater than 20%, 10%, and 5% of the normalization factor (100 Hz). The average pitch frequencies for the Male1 and Male2 speakers were 186.95 Hz and 213.82 Hz respectively.

Speaker	**Percentage of Pitch Errors greater than x% of 100 Hz**		
	>20%	**>10%**	**>5%**
Female1	25.61	34.39	41.93
Female2	22.18	29.92	34.10
Combined	24.05	32.35	38.36

Table 4.19: Percentage of pitch errors, for a male/female mixture (SIR = -6 dB), which were greater than 20%, 10%, and 5% of the normalization factor (100 Hz). The average pitch frequencies for the male and female speakers were 109.17 Hz and 186.44 Hz respectively.

Speaker	**Percentage of Pitch Errors greater than x% of 100 Hz**		
	>20%	**>10%**	**>5%**
Male	21.92	29.28	33.33
Female	15.04	18.16	20.99
Combined	18.03	23.00	26.36

4.3 Constrained Nonlinear Optimization

Our constrained nonlinear least squared optimization algorithm is used to separate overlapping voiced speech. We tested this algorithm using simulated speech data and from a single frame of co-channel speech data taken from the TIMIT_COC database. The simulated data consisted of two vocalic speech segments, represented as a sum of sine waves. The co-channel speech data was taken from a frame of overlapping vocalic speech produced from a male and female talker. Our results are presented below.

The first scenario was to test the optimization algorithm on data comprised of two simulated vocalic speech segments, represented as a sum of sine waves where the unknown variables were the amplitude, phase, and frequency of two harmonics. The sum of these two simulated speech segments was used to model a segment of co-channel speech. This signal can be represented as

$$\boldsymbol{s}(\boldsymbol{x},\boldsymbol{n}) = \sum_{i=1}^{4} \boldsymbol{x}_1[i] cos(2\pi n \boldsymbol{x}_2[i] + \boldsymbol{x}_3[i]) \qquad (4.1)$$

where $\boldsymbol{x}_j$ is a set of vectors representing the amplitude ($\boldsymbol{x}_1$), frequency ($\boldsymbol{x}_2$), and phase ($\boldsymbol{x}_3$) for each harmonic. The reference values for our simulated co-channel speech segment were

$$\begin{aligned} \boldsymbol{x}_1 &= [5 \quad 7 \quad 10 \quad 8]^t \\ \boldsymbol{x}_2 &= [101 \quad 150 \quad 205 \quad 303]^t \\ \boldsymbol{x}_3 &= [.1 \quad -.2 \quad -.3 \quad -.3]^t \end{aligned} \qquad (4.2)$$

where the first and third harmonics were associated with our first signal and the second and fourth harmonics were associated with our second signal. The initial conditions of our unknown variables were

$$\begin{aligned} \boldsymbol{x}_1 &= [8 \quad 9 \quad 4 \quad 5]^T \\ \boldsymbol{x}_2 &= [98 \quad 140 \quad 196 \quad 310]^T \\ \boldsymbol{x}_3 &= [.15 \quad -.24 \quad -.33 \quad -.32]^T \end{aligned} \tag{4.3}$$

Constrained nonlinear optimization was applied to the simulated co-channel signal to minimize the squared error between the reference co-channel signal and the estimated co-channel signal by optimizing the amplitude, phase, and center frequency parameters. The input values to the optimization routine were the initial conditions of our unknown variables. The reference signal was generated using the reference values. The signals were Hanning weighted prior to optimization. The solution converged in 47 iterations with a least squared error of $2.9189e^{-7}$. The final values of our variables were

$$\begin{aligned} \boldsymbol{x}_1 &= [4.9637 \quad 6.8817 \quad 9.6020 \quad 7.6442]^T \\ \boldsymbol{x}_2 &= [100.8143 \quad 149.9649 \quad 204.9410 \quad 303.0004]^T \\ \boldsymbol{x}_3 &= [.5978 \quad -1.3418 \quad -2.9090 \quad -2.3649]^T \end{aligned} \tag{4.4}$$

A comparison of the two simulated signals, their initial conditions and the resulting signal after optimization are given in Figure 4.9 and Figure 4.10.

In our next experiment, we tested the ability of the constrained nonlinear optimization algorithm to estimate the unknown harmonic parameters of two vocalic speech segments taken from the TIMIT_COC database. The two test signals are shown in Figure 4.11 (a) and (b) with the co-channel signal (the sum of both signals) shown in

Figure 4.11 (c). The speech signals were summed on the computer at an average SIR = 0 dB. The speech segment in Figure 4.11 (a) was taken from the vocalic '*ao*' sound produced from a male talker when saying the word 'bought'. The signal in Figure 4.11 (b) was taken from the vocalic '*ay*' sound as would be produced from a female talker when saying the word 'bite'. These fundamental frequencies of these two speech segments are 130 Hz and 245 Hz respectively.

Our speech segment, represented as a sum of sine waves each with a given amplitude, frequency and phase, can be written as

$$s[n] = \sum_{k=1}^{M} a_k cos(\omega_k n + \phi_k) \tag{4.5}$$

This can be rewritten as

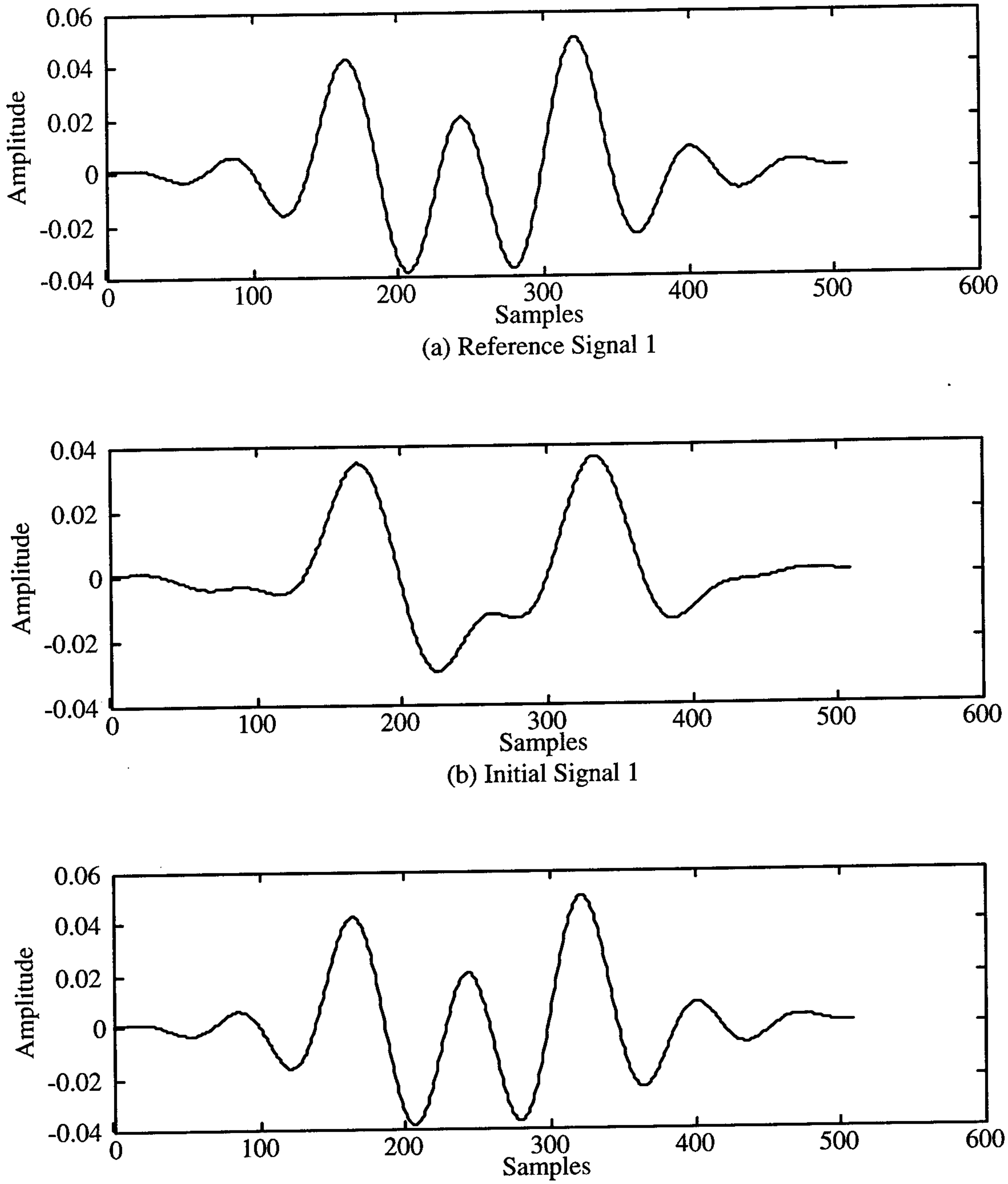
$$s[n] = \sum_{k=1}^{M} a_k cos(\omega_k n)cos(\phi_k) - a_k sin(\omega_k n)sin(\phi_k) \tag{4.6}$$

$$s[n] = \sum_{k=1}^{M} \alpha_k cos(\omega_k n) + \beta_k sin(\omega_k n) \tag{4.7}$$

where

$$\begin{aligned} \alpha_k &= a_k cos(\phi_k) \\ \beta_k &= -a_k sin(\phi_k) \end{aligned} \tag{4.8}$$

The unknown variables then become α_k, β_k and ω_k..

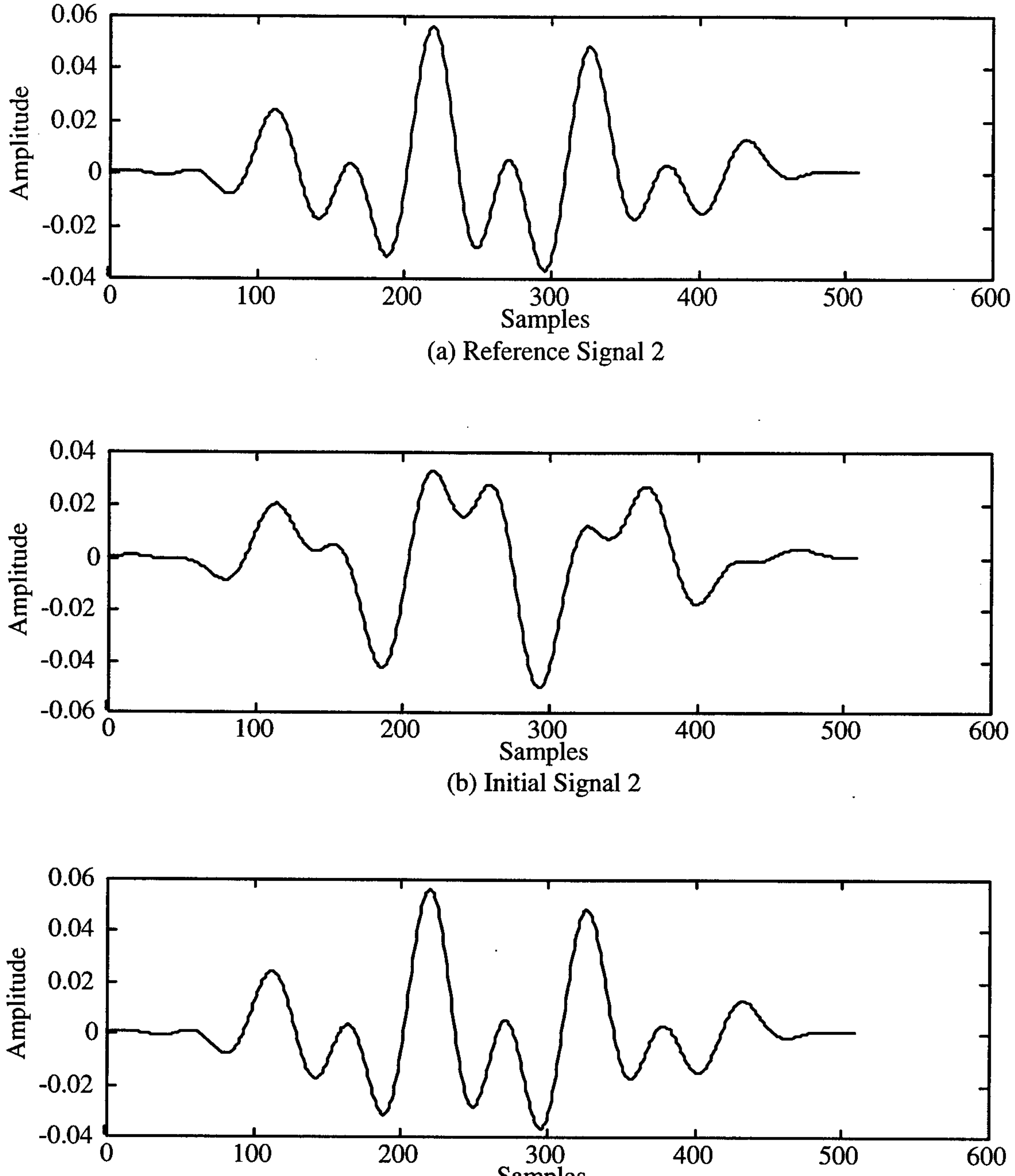


(a) Reference Signal 1

(b) Initial Signal 1

(c) Reconstructed Signal 1

Figure 4.9: Reconstruction of simulated vocalic speech signal using constrained nonlinear optimization, comparing (a) reference signal 1, (b) initial signal 1, and (c) reconstructed signal 1.



(a) Reference Signal 2

(b) Initial Signal 2

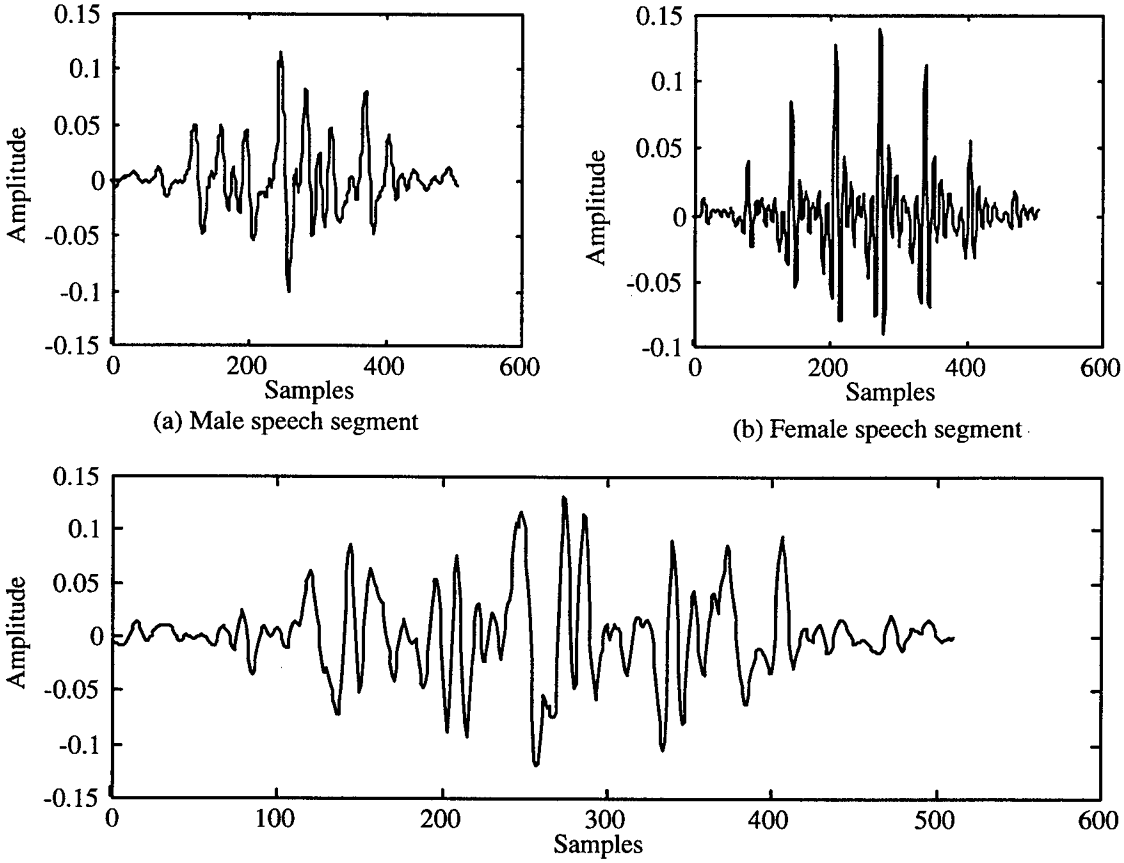
(c) Reconstructed Signal 2

Figure 4.10: Reconstruction of simulated vocalic speech signal using constrained nonlinear optimization, comparing (a) reference signal 2, (b) initial signal 2, and (c) reconstructed signal 2.

The initial conditions for this test are similar to those in the previous test. The harmonics of the combined speech signal were set to integer multiples of the fundamental frequencies of both speech segments. The amplitudes were set to a nominal value (0.10). The upper and lower bounds on the frequencies were +10 Hz and -10 Hz respectively. The upper and lower bounds on the amplitude terms were set to +1 and -1 respectively. We selected the first eight harmonics of each segment for optimization. This represented a total of 48 unknown variables.

Results of our test are presented in Figure 4.12 - 4.15. The optimization routine converged with an overall least squared error of 4.65×10^{-2}. This error represents approximately 5.9% of the overall average energy in the co-channel signal. Specifically, the percentage of error, between the original male speech segment and the reconstructed speech segment was 2.4% while that for the female speech segment was 8.67%.

Figure 4.12 shows the original Hanning weighted male speech segment in (a) and the recovered speech segment in (b). Figure 4.13 displays the magnitude frequency spectrum of the original male speech segment in (a) and the recovered speech segment in (b). Figure 4.14 shows the original Hanning weighted female speech segment in (a) and the recovered speech segment in (b), while Figure 4.15 displays the magnitude frequency spectrum of the female original speech segment in (a) and the recovered speech segment in (b). From these plots we see that the recovered speech segments for both speakers accurately resemble their original speech segments.

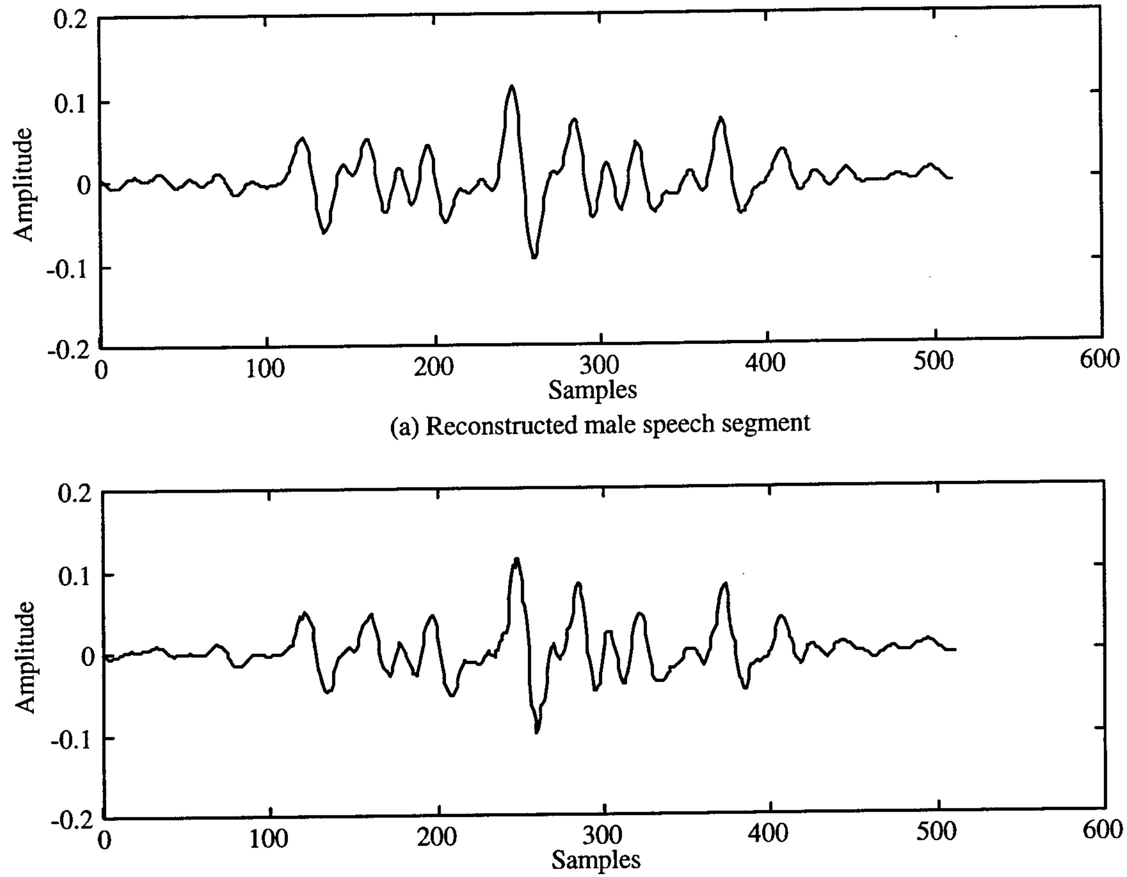


(a) Male speech segment

(b) Female speech segment

(c) Co-channel speech segment

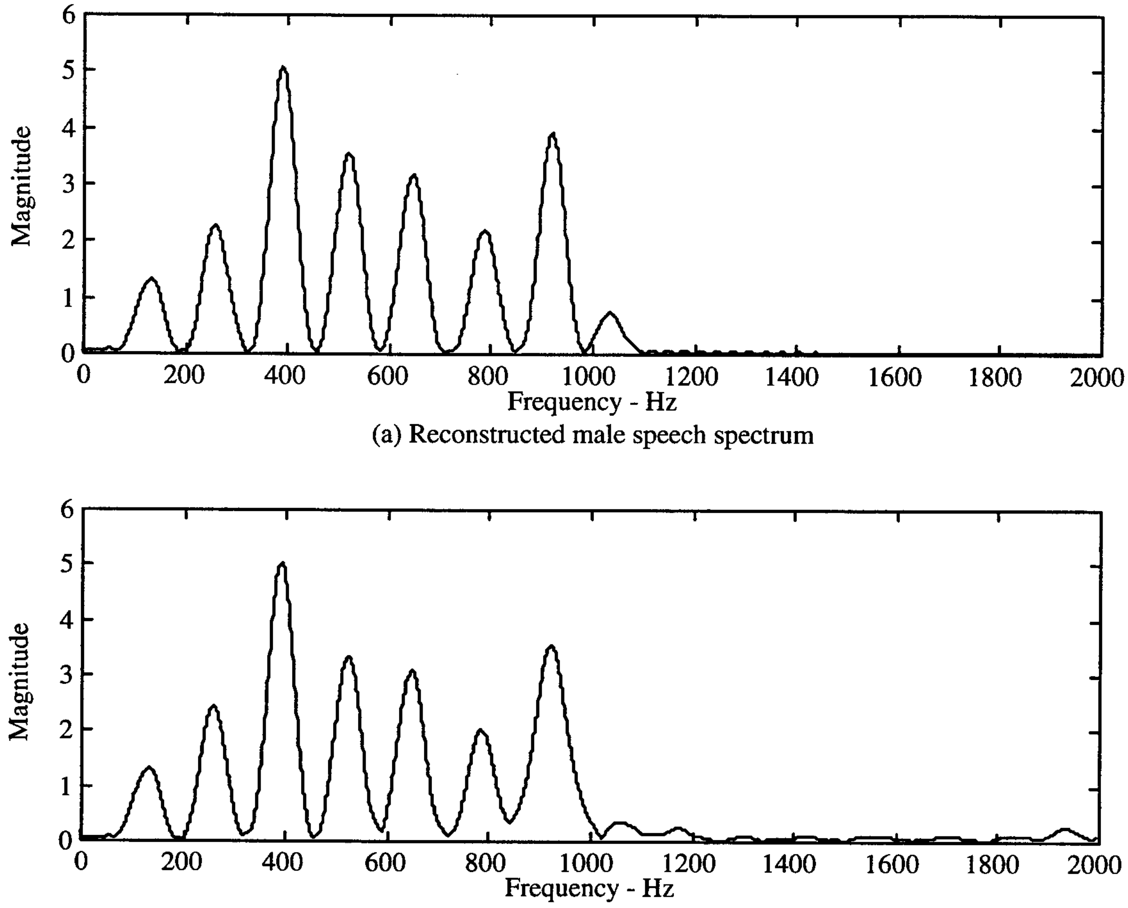
Figure 4.11: Hanning weighted speech segments, (a) male speaker, (b) female speaker, and (c) co-channel speech (male+female).



(a) Reconstructed male speech segment

(b) Original male speech segment

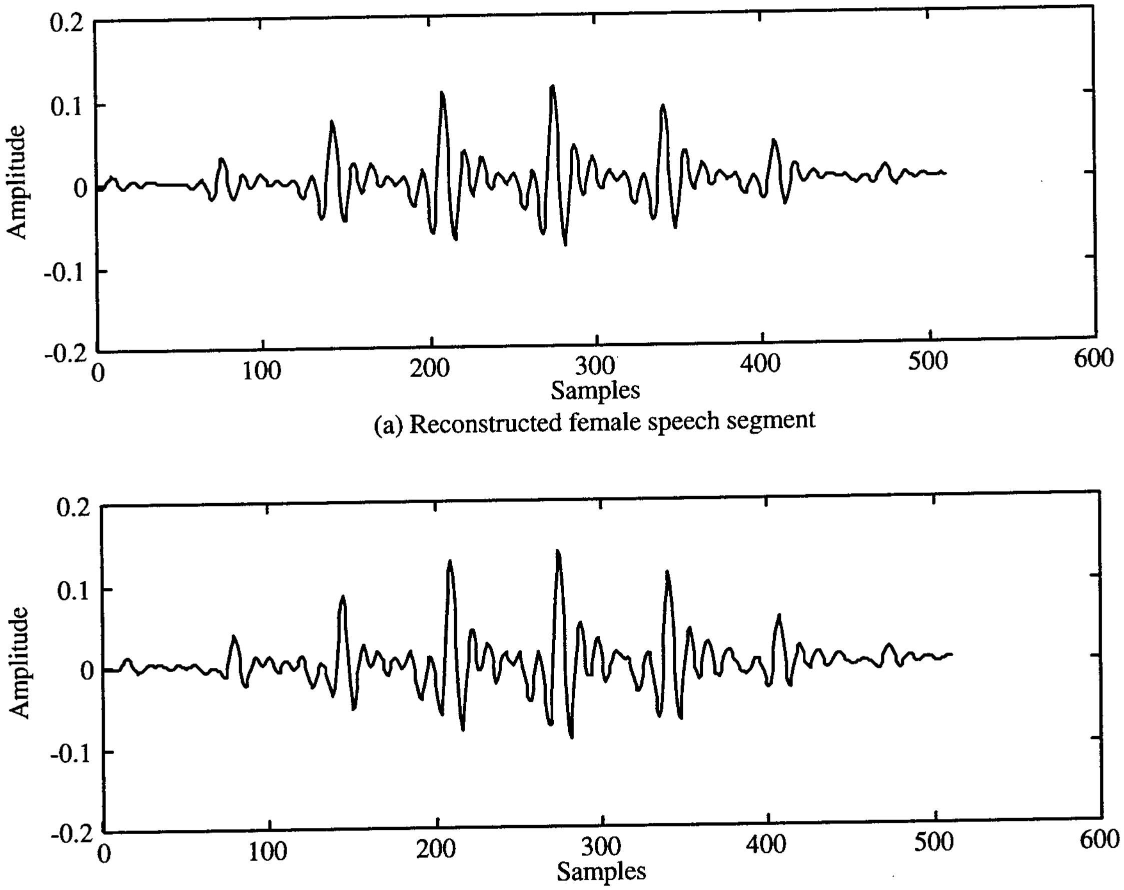
Figure 4.12: Male speech segment using constrained nonlinear optimization to separate co-channel speech, (a) reconstructed speech segment, (b) original speech segment.



(a) Reconstructed male speech spectrum

(b) Original male speeech spectrum

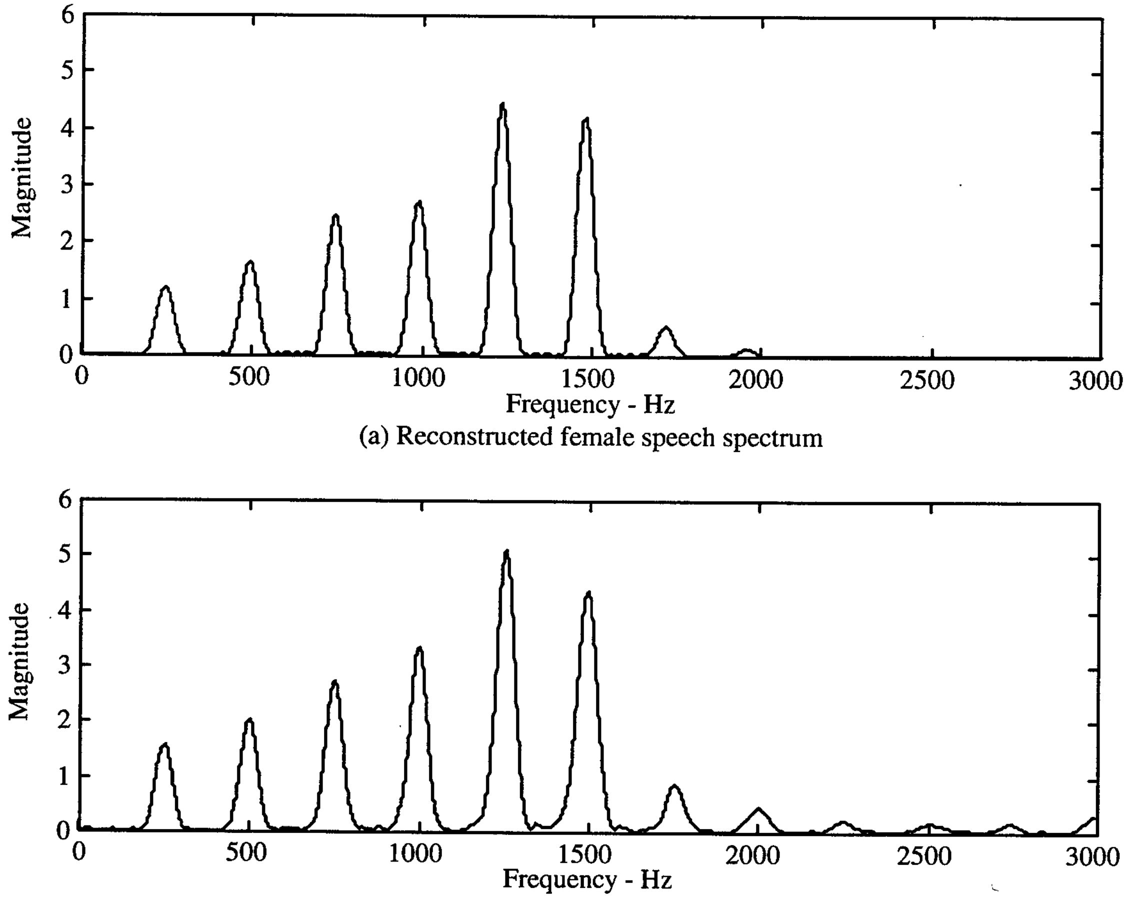
Figure 4.13: Magnitude spectrum of male speech segment using constrained nonlinear optimization to separate co-channel speech, (a) reconstructed spectrum, (b) original spectrum.



(a) Reconstructed female speech segment

(b) Original female speech segment

Figure 4.14: Female speech segment using constrained nonlinear optimization to separate co-channel speech, (a) reconstructed speech segment, (b) original speech segment.



(a) Reconstructed female speech spectrum

(b) Original female speech spectrum

Figure 4.15: Magnitude spectrum of female speech segment using constrained nonlinear optimization to separate co-channel speech, (a) reconstructed spectrum, (b) original spectrum.

4.4 Simultaneous Adaptive Co-Channel Speaker Separation

The main goal of our research was to separate the speech of two talkers recorded over a single channel. An effective speaker separation system must operate autonomously, that is, it must process the co-channel speech signal without *a priori* information (i.e. pitch contours). Our co-channel speaker separation system is an autonomous system which performs the following tasks:

1. Train the VSDA, using supervised learning, to classify the co-channel speech.
2. Create co-channel speech data at a given SIR.
3. Calculate the co-channel voicing state and pitch contour for each speaker.
4. Apply the proper processing technique, based on the voicing state of each speaker, to separate the speech into desired and interfering speech segments
5. Reconstruct the two speech signals from the processed speech segments.

We have tested our speaker separation system, presented in Section 3.1.3, on male/female, male/male and female/female speech mixtures using speech signals from the TIMIT_COC database, outlined in Tables 4.3 - 4.5. The data consisted of the same four, two speaker co-channel speech sentences used to test the VSDA and the joint pitch estimation algorithm. The Bayes classifier was used in the VSDA for these experiments. The optimization routine was simplified by excluding the center frequencies from the optimization routine to reduce the overall processing time. This reduced the processing time by a factor of 20. The phase and amplitude were estimated using the optimization routine, while the center frequency terms were estimated using our harmonic selection algorithm presented in Section 3.6. Cumulative results of the four co-channel speech sentences for each speech mixture are presented below. We also conducted tests using data taken from the RPI_COC speech database. Results on a simulated co-channel speech signal and a real co-channel speech signal are also presented below.

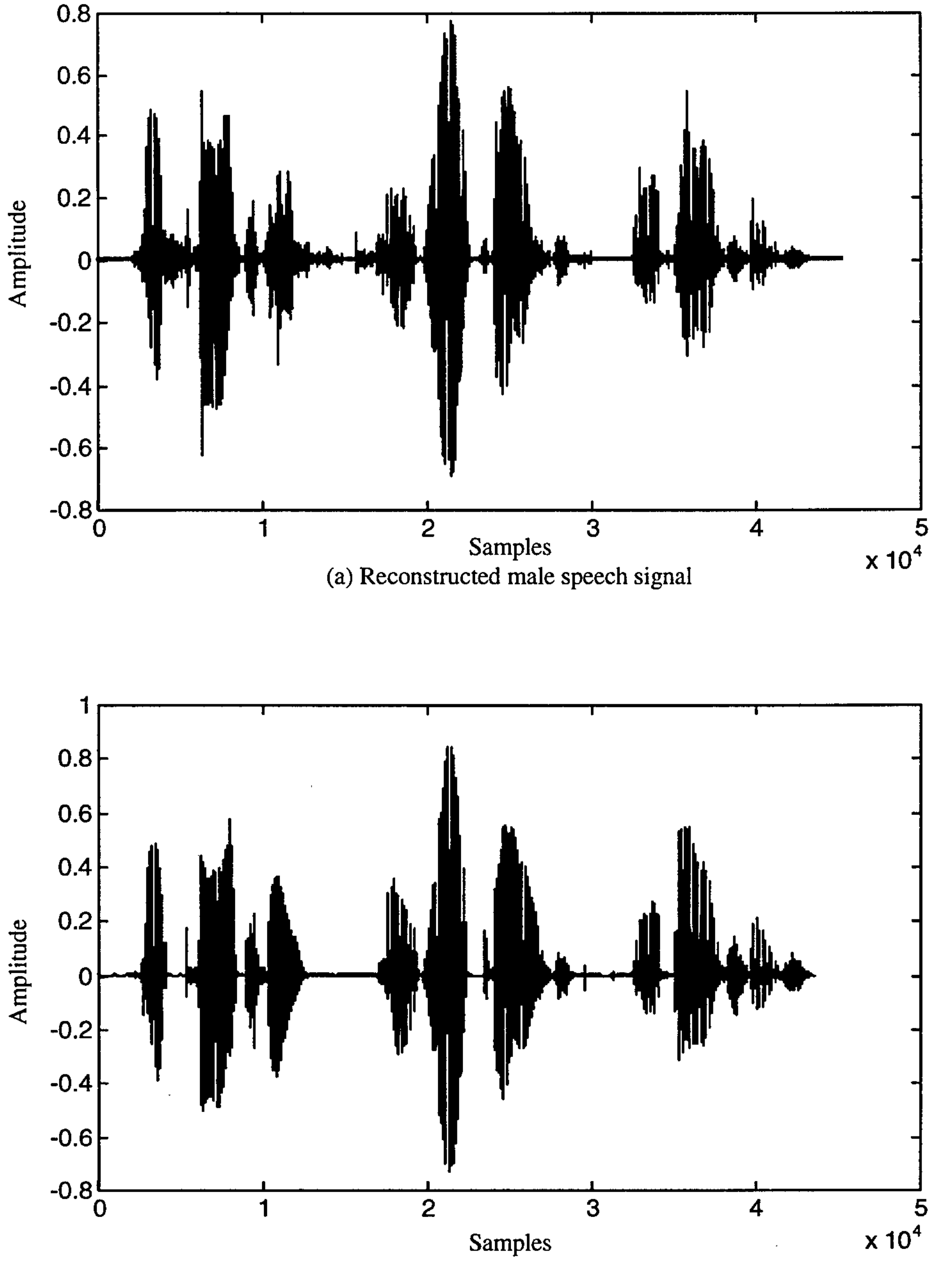
To compensate for the performance of the voicing state determination algorithm, we introduced a decision structure into the voiced/voiced branch of the system. Our VSDA may erroneously classify mixed voiced speech as all voiced speech. In our system we implemented a method which would first estimate the stronger speech signal, based on the pitch frequency, using harmonic selection. Then an estimate of the weaker signal using two different methods was made. The first method, assuming the weaker signal was voiced, estimated the signal using harmonic magnitude suppression. The second method assumed the weaker signal was unvoiced and estimated the signal using a highpass filter. Two co-channel signals were created, the first was sum of the estimate of the stronger signal plus the estimate of the assumed voiced signal, and the second was created from the estimate of the stronger signal and the estimate of the assumed unvoiced signal. These two signals were then compared to the measured co-channel signal. The co-channel signal, which had the smallest error (see equation 4.9) would decide the correct co-channel voicing state and separation would be performed as described in Chapter 3.

Figure 4.16 and Figure 4.17 provides an example showing a comparison between two speech signals before mixing and after separation for a male/female speech mixture at SIR = 0 dB. The original and reconstructed speech signals for the male and female talker are shown in Figure 4.16 and Figure 4.17 respectively. The reconstructed speech signals are intelligible and sound similar to the original uncorrupted speech signals.

Tables 4.20 - 4.23 provides a relative error measurements between the original speech signal and the reconstructed speech signal for both talkers and for the co-channel speech signals in the ten, two-speaker co-channel speech combinations. We define this error measurement as

$$Relative\_Error = \frac{\sum_{i=1}^{N}\left[s[i]-\hat{s}[i]\right]^2}{\sum_{i=1}^{N}\left[s[i]\right]^2} \tag{4.9}$$

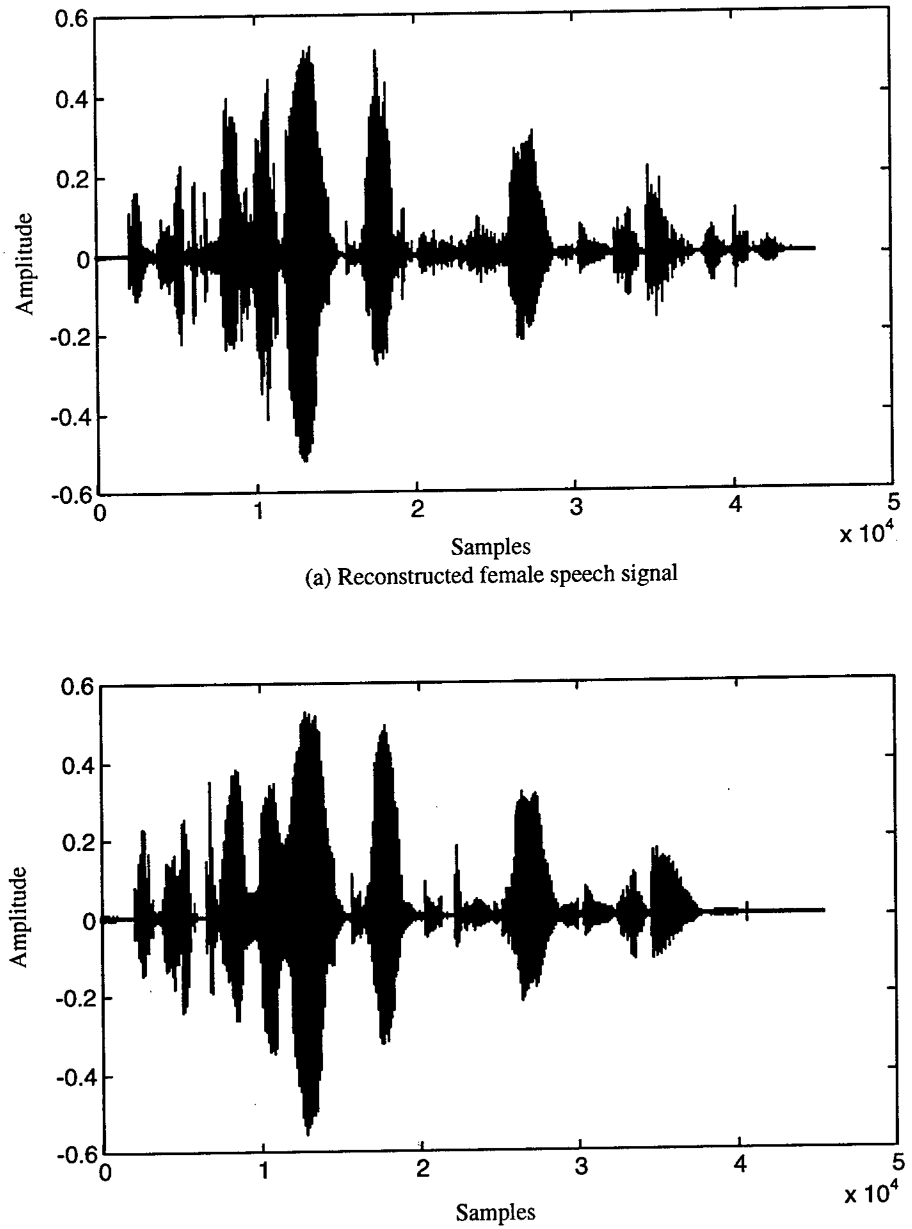
where N is the total length of the speech signal. Results are presented for each speech sentence pair for all speech mixtures. The last column in each table represents the average of the relative error measurements over the ten sentence combinations for each talker. While these values do not give an indication of the intelligibility of the speech, they do provide a relative error measure across mixtures. We can see from these results that our speaker separation system performed better on male/female co-channel speech signals than on same gender co-channel speech segments. These results consistent with the previous two experiments.



(a) Reconstructed male speech signal

(b) Original male speech signal

Figure 4.16: Male speech signal, from a male/female co-channel speech mixture at SIR = 0 dB. (a) Reconstructed male speech signal and (b) original male speech signal.



(a) Reconstructed female speech signal

(b) Original female speech signal

Figure 4.17: Female speech signal, from a male/female co-channel speech mixture at SIR = 0 dB. (a) Reconstructed female speech signal and (b) original female speech signal.

Table 4.20: Error measurements between reconstructed signal and original signal in a male/female mixture with SIR = 0 dB. Mean represents mean error over 4 sentence combinations for male (M), female (F), and Co-channel (C) signals.

M/F Mixture 0 dB	**Co-Channel Speech Sentence Relative Error**				
	C1	**C2**	**C3**	**C4**	**Mean**
M	.6148	.4551	.5817	.5902	.5604
F	.5190	.4345	.3238	.4408	.4295
C	.2811	.2438	.2350	.2636	.2559

Table 4.21: Error measurements between reconstructed signal and original signal in a male/male mixture with SIR = 0 dB. Mean represents mean error over 4 sentence combinations for Male1 (M1), Male2 (M2), and Co-channel (C) signals.

M/M Mixture 0 dB	**Co-Channel Speech Sentence Relative Error**				
	C1	**C2**	**C3**	**C4**	**Mean**
M1	.9550	.7516	.8854	.9646	.8891
M2	.6231	.5621	.7064	.7719	.6659
C	.2960	.3139	.2734	.2547	.2845

Table 4.22: Error measurements between reconstructed signal and original signal in a female/female mixture with SIR = 0 dB. Mean represents mean error over 4 sentence combinations for Female1 (F1), Female2 (F2), and Co-channel (C) signals.

F/F Mixture 0 dB	**Co-Channel Speech Sentence Relative Error**				
	C1	**C2**	**C3**	**C4**	**Mean**
F1	.9116	.7866	.8723	.8325	.8508
F2	1.2075	1.0048	1.1318	1.3880	1.1830
C	.2574	.2193	.2178	.2202	.2287

Table 4.23: Error measurements between reconstructed signal and original signal in a male/female mixture with SIR = -6 dB. Mean represents mean error over 4 sentence combinations for male (M), female (F), and Co-channel (C) signals.

M/F Mixture -6 dB	**Co-Channel Speech Sentence Relative Error**				
	C1	**C2**	**C3**	**C4**	**Mean**
M	.7688	.6095	.8618	.8628	.7757
F	.3570	.3428	.2821	.3704	.3381
C	.2208	.2231	.2182	.2344	.2242

To determine the impact of excluding the center frequency terms as optimization parameters, we repeated the male/female, SIR = 0 dB experiment with the center frequency terms included in the optimization. The average relative error, for the male, female, and co-channel signals dropped by 1.29%, 1.05%, and 35.15%, respectively when the center frequency terms were included.

To test the effectiveness of our speaker separation system, we have shown in Figure 4.18 through Figure 4.21 the change of the average relative error, due to the introduction of *a priori* information, for the three different speech mixtures (male/female, male/male and female/female). The labeling along the x-axis represents those results obtained when there was no cheating (NC), when the true voicing states (TVS) were used instead of the estimated voicing states, when the true pitch contours (TP) were used instead of the measured pitch contours, and when both the true voicing states and true pitch contours (TVS & TP) were used.

In Figure 4.18 we discover that our system design is optimized to separate two speakers of a different gender. Figure 4.19 through Figure 4.21 demonstrate the relative importance to our system design in obtaining accurate pitch contours when the two speakers were of the same gender. When both speakers were of the same gender, the relative error dropped significantly when the true pitch contours were used. The pitch contours are crucial to providing an accurate estimate of each talker's harmonic peaks in the co-channel speech.

These graphs also show that an accurate voicing state determination algorithm is less important. Our system design is somewhat resilient to errors in estimating the voicing state. Our voicing state determination algorithm performed between 70% and 80% overall correct classification. This is sufficient. We can verify this by referring back to Table 4.6 - Table 4.15. Here we see that a major portion of these errors associated with misclassification of the voicing state occurs when mixed voiced speech (V/UV and UV/V) is labeled as all voiced speech (V/V). Our separation method that is used to separate V/V speech mixtures relies on the pitch estimates to perform harmonic selection (Section 3.6). This method obtains an estimate of the stronger harmonics first and then estimates the weaker harmonics. For the case when the speech segments true voicing state is V/UV or UV/V and is misclassified as V/V, we would obtain an accurate estimate of the voiced speech segment and a poor estimate of the unvoiced speech segment. However, as stated previously, we can replace an unvoiced sound with a noise-like sound and still have intelligible speech [45]. The source of error occurs when, based on the two pitch estimates, two harmonics overlap and the optimization routine places energy at this harmonic location into the unvoiced segment of speech. This erroneously transfers energy to the unvoiced speaker.

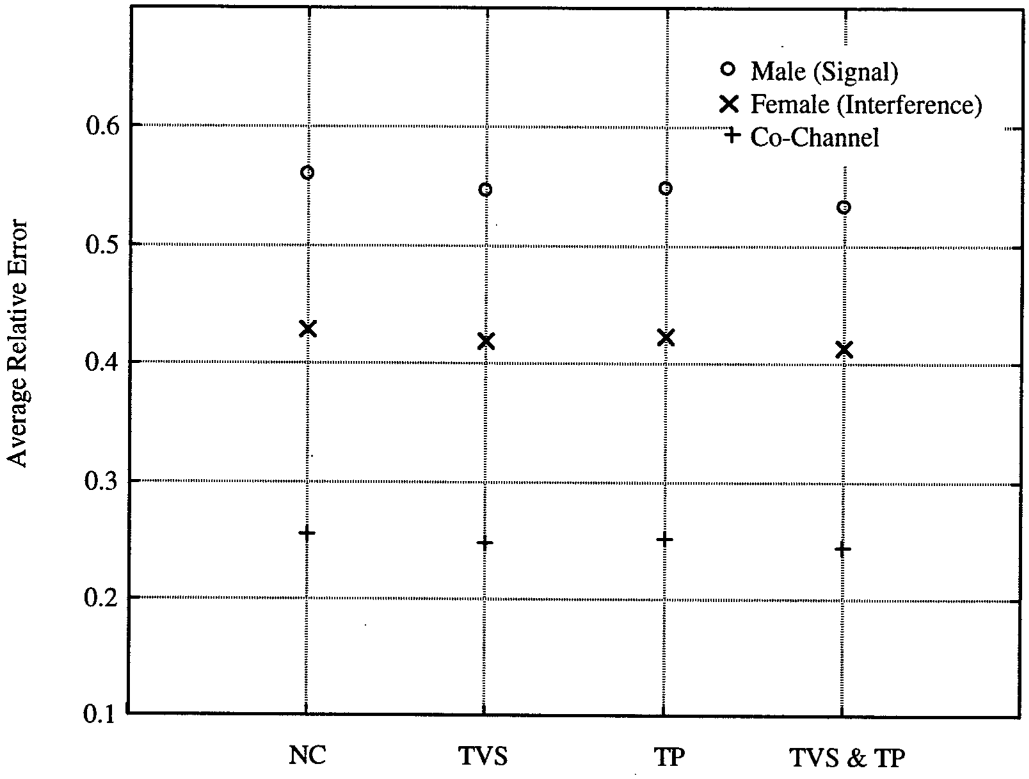


Figure 4.18: Plot of the average relative error versus increasing a priori information, for male/female speech mixtures at SIR = 0 dB. No Cheating (NC); True Voicing State (TVS); True Pitch (TP).

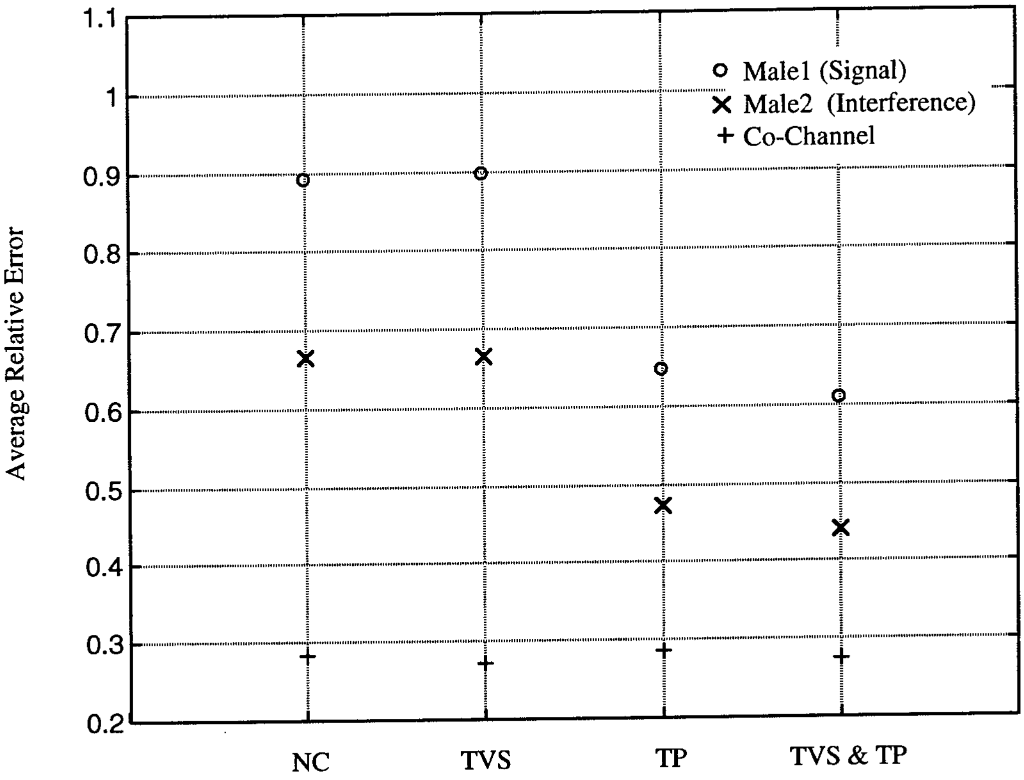


Figure 4.19: Plot of the average relative error versus increasing a priori information, for male/male speech mixtures at SIR = 0 dB. No Cheating (NC); True Voicing State (TVS); True Pitch (TP).

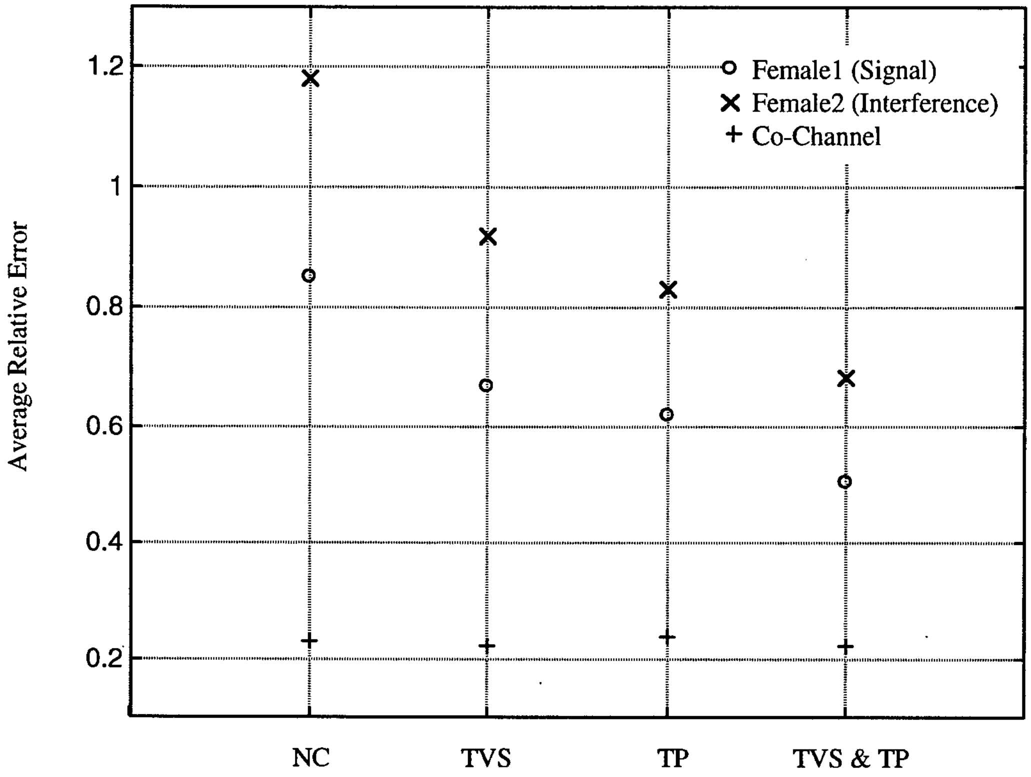


Figure 4.20: Plot of the average relative error versus increasing a priori information, for female/female speech mixtures at SIR = 0 dB. No Cheating (NC); True Voicing State (TVS); True Pitch (TP).

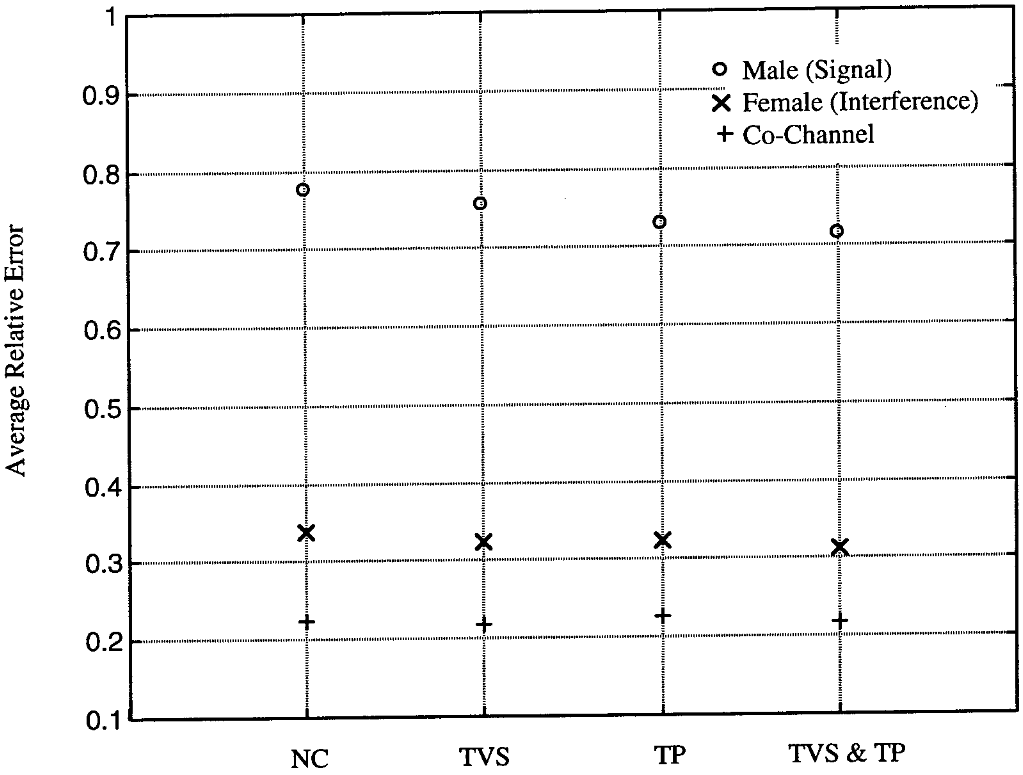


Figure 4.21: Plot of the average relative error versus increasing a priori information, for male/female speech mixtures at SIR = -6 dB. No Cheating (NC); True Voicing State (TVS); True Pitch (TP).

The error term between our original signal and reconstructed signal can be broken down into errors due to voicing state misclassification (Type I) and errors due to processing (Type II). Type I errors occur when co-channel voicing state of a segment of speech is incorrect. Type II errors occur when the co-channel voicing state is labeled correctly, but the separation process performs poorly in separating the two speech signals. We can further consider those errors associated with each co-channel voicing state. Referring to Table 4.24 - Table 4.29, we present cumulative relative error measurements for each speaker in the four sentence pairs of male/female, male/male and female/female speech mixtures. In these tables, we have identified the percentage of frames in which the error measurement exceeded a given threshold. Along the first row in the table is the percentage of frames in which the error measurement exceeded our given threshold for each co-channel voicing state. The second row represents the percentage of frames, for each co-channel voicing state, which exceeded the threshold but whose voicing state was correctly classified. The third row represents our Type I error, or the percentage of the total error measurement that resulted from misclassification of the voicing state. The last row represents our Type II error, the percentage of the total error measurement that resulted from our separation processing.

Table 4.24: Contribution and percentage of errors per voicing state, for the Male in Male/Female speech mixtures at SIR = 0 dB.

Male	SIL	V/V	V/UV	UV/V	UV/UV
% of frames with errors > threshold	12.09	8.38	0.78	44.25	34.50
% of frames with errors > threshold & correct voicing state	12.09	6.82	0	39.57	29.43
Type I error (VSD)	0	0.08	0.08	3.05	5.07
Type II error (Processing)	1.98	0.32	0	40.02	49.40

Table 4.25: Contribution and percentage of errors per voicing state, for the Female in Male/Female speech mixtures at SIR = 0 dB.

Female	SIL	V/V	V/UV	UV/V	UV/UV
% of frames with errors > threshold	11.76	24.61	37.43	2.67	23.53
% of frames with errors > threshold & correct voicing state	4.29	6.04	7.60	0.39	6.24
Type I error (VSD)	0	2.04	25.10	0.60	4.15
Type II error (Processing)	7.53	4.13	25.16	0.26	31.03

Table 4.26: Contribution and percentage of errors per voicing state, for the Male1 in Male1/Male2 speech mixtures at SIR = 0 dB.

Male1	SIL	V/V	V/UV	UV/V	UV/UV
% of frames with errors > threshold	17.06	13.21	5.04	33.98	30.71
% of frames with errors > threshold & correct voicing state	17.06	10.09	0.74	23.15	24.04
Type I error (VSD)	0	0.10	0.17	22.73	13.60
Type II error (Processing)	2.94	0.25	0.01	34.82	25.38

Table 4.27: Contribution and percentage of errors per voicing state, for the Male2 in Male1/Male2 speech mixtures at SIR = 0 dB.

Male2	SIL	V/V	V/UV	UV/V	UV/UV
% of frames with errors > threshold	0.47	59.53	22.79	4.65	12.56
% of frames with errors > threshold & correct voicing state	0.15	15.88	3.71	0	2.82
Type I error (VSD)	0	3.35	31.77	2.22	1.38
Type II error (processing)	0.12	36.46	10.68	0	14.02

Table 4.28: Contribution and percentage of errors per voicing state, for the Female1 in Female1/Female2 speech mixtures at SIR = 0 dB.

Female1	**SIL**	**V/V**	**V/UV**	**UV/V**	**UV/UV**
% of frames with errors > threshold	1.22	29.67	11.38	35.37	22.36
% of frames with errors > threshold & correct voicing state	1.22	21.95	0.81	12.20	15.45
Type I error (VSD)	0	3.20	3.37	27.60	2.47
Type II error (Processing)	0.30	5.42	0.20	7.29	50.15

Table 4.29: Contribution and percentage of errors per voicing state, for the Female2 in Female1/Female2 speech mixtures at SIR = 0 dB.

Female2	**SIL**	**V/V**	**V/UV**	**UV/V**	**UV/UV**
% of frames with errors > threshold	11.99	25.50	34.80	5.24	22.47
% of frames with errors > threshold & correct voicing state	28.46	49.59	44.72	1.63	47.15
Type I error (VSD)	0	0.07	61.84	0.03	3.45
Type II error (Processing)	0.78	0.49	18.57	0.02	14.75

From these measurements we see that a small number of the V/V frames which were correctly classified, had error measurements above the threshold but that the error due to processing was consistently low for all the speech mixtures. The major portion of this error is a result of the harmonics between two speakers overlapping. A high percentage of error was associated with the misclassification of mixed voiced (V/UV, UV/V) and UV/UV speech segments. A large portion of this error occurs when one speaker is silent and the other is voiced or unvoiced. In these cases, the silent segment is replaced with an unvoiced speech segment. This error term accumulates when one speaker is silent for extended intervals of time. Specifically, for the male/female speech mixtures, the male speech signals were consistently shorter in length than the female speech signals, resulting in a large percentage of error due to processing for the UV/V and UV/UV states. For the same gender speech mixtures, a large portion of the total error measurement was due to voicing state error in the V/UV and UV/V co-channel speech states. This is a result of the mixed speech being incorrectly identified as V/V speech.

Our next experiment was to test the overall effectiveness of our speaker separation system when the signal to interferer ratio (SIR) varied. We conducted experiments at average SIRs of 0, -6 and -12 dB. These tests were conducted on the same set of speech signals used in the previous test. Our results are presented in Figure 4.22 through Figure 4.28.

In Figure 4.22 through Figure 4.24 we show the average relative error for the three different speech mixtures versus a decreasing signal to interferer ratio (SIR). Here we see that for all three cases the error measurement between the original signal and the reconstructed signal decreased for the stronger speaker and increased for the weaker speaker. The co-channel error measurement stayed relatively constant.

As we decrease the SIR (increase the energy of the interfering signal), we see that for the voicing state determination algorithm, the overall percentage of correct detects decreased for the male/female and male/male speech mixtures, but increased for the female/female speech mixture. See Figure 4.25. In Figure 4.26 we see that the percentage of correct detects for each voicing state varied only slightly as the SIR decreased for the male/female case. For the same gender case, the performance of the VSDA for each co-channel voicing state stayed relatively constant except for the mixed voiced states (V/UV and UV/V). In this case the percentage of correct detects for the UV/V case decreased significantly as the SIR decreased. This is a result of the voiced speech from the stronger speaker masking the unvoiced speech from the weaker speaker. In these cases the speech segment is misclassified as V/UV or V/V speech. This variation was not as significant for the male/female speech mixtures.

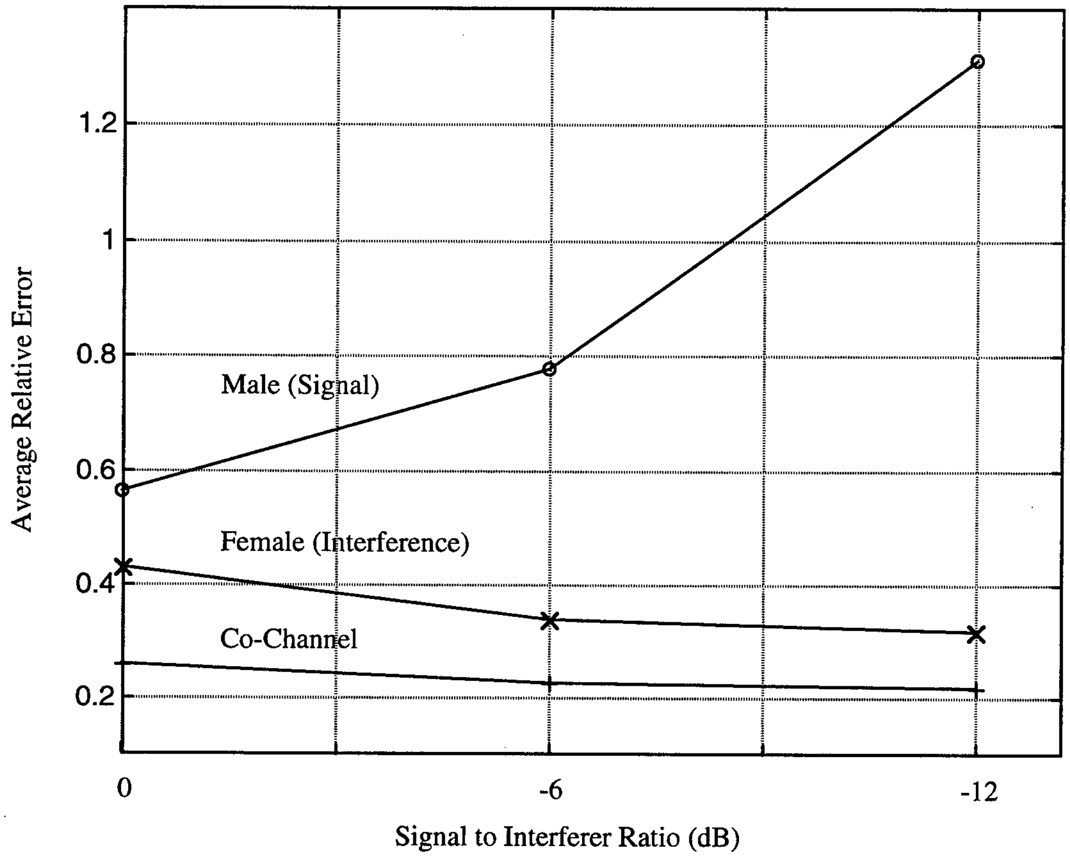


Figure 4.22: Plot of the average relative error versus SIR = 0, -6 and -12 dB for male/female speech mixtures.

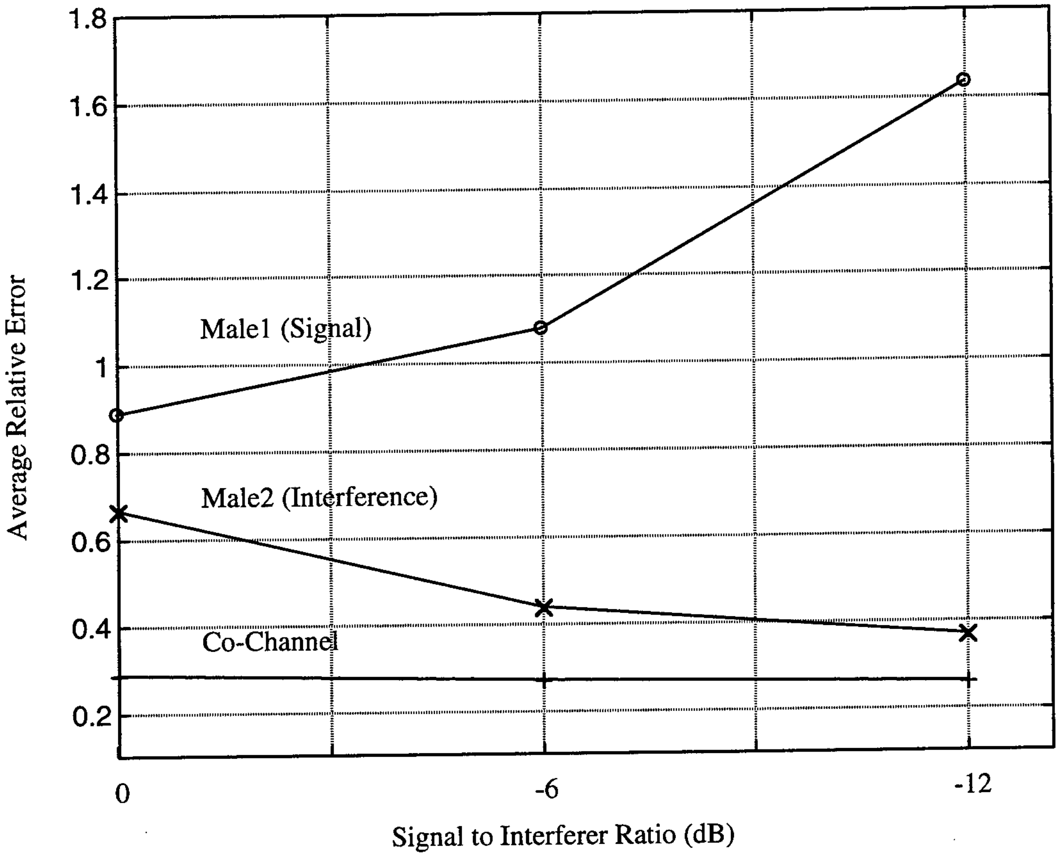


Figure 4.23: Plot of the average relative error versus SIR = 0, -6, and -12 dB for male/male speech mixtures.

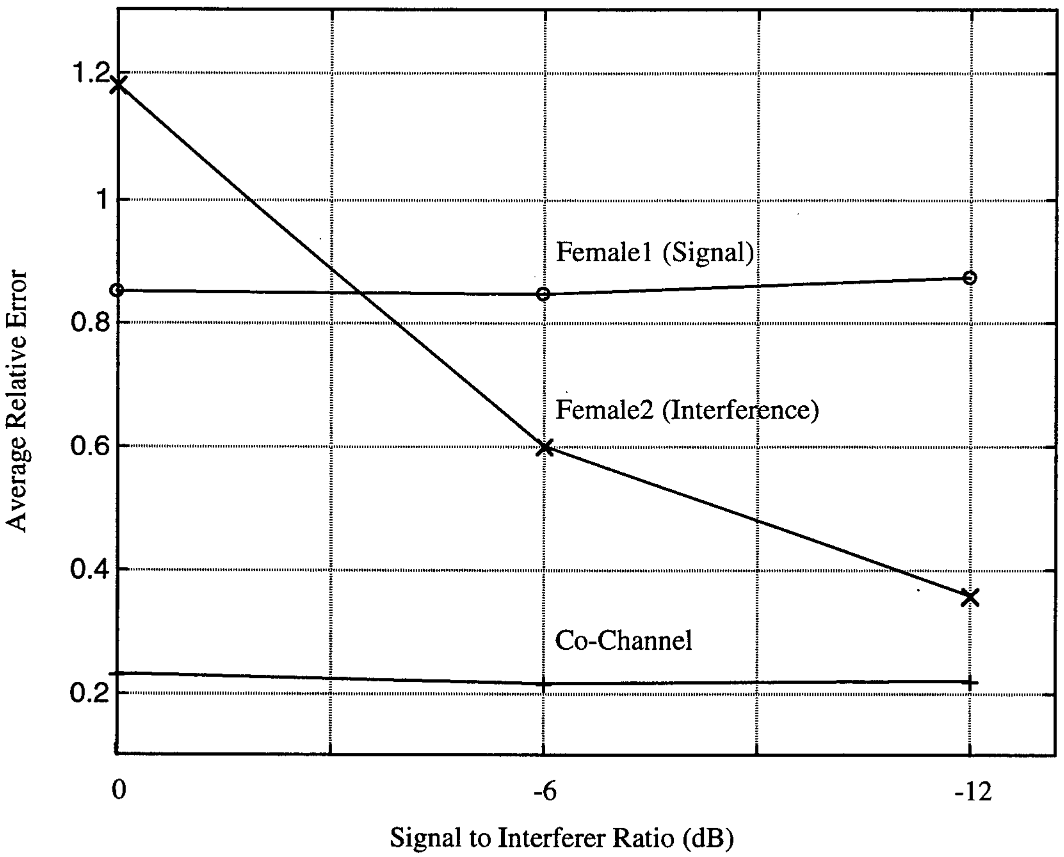


Figure 4.24: Plot of the average relative error versus SIR = 0, -6, and -12 dB for female/female speech mixtures.

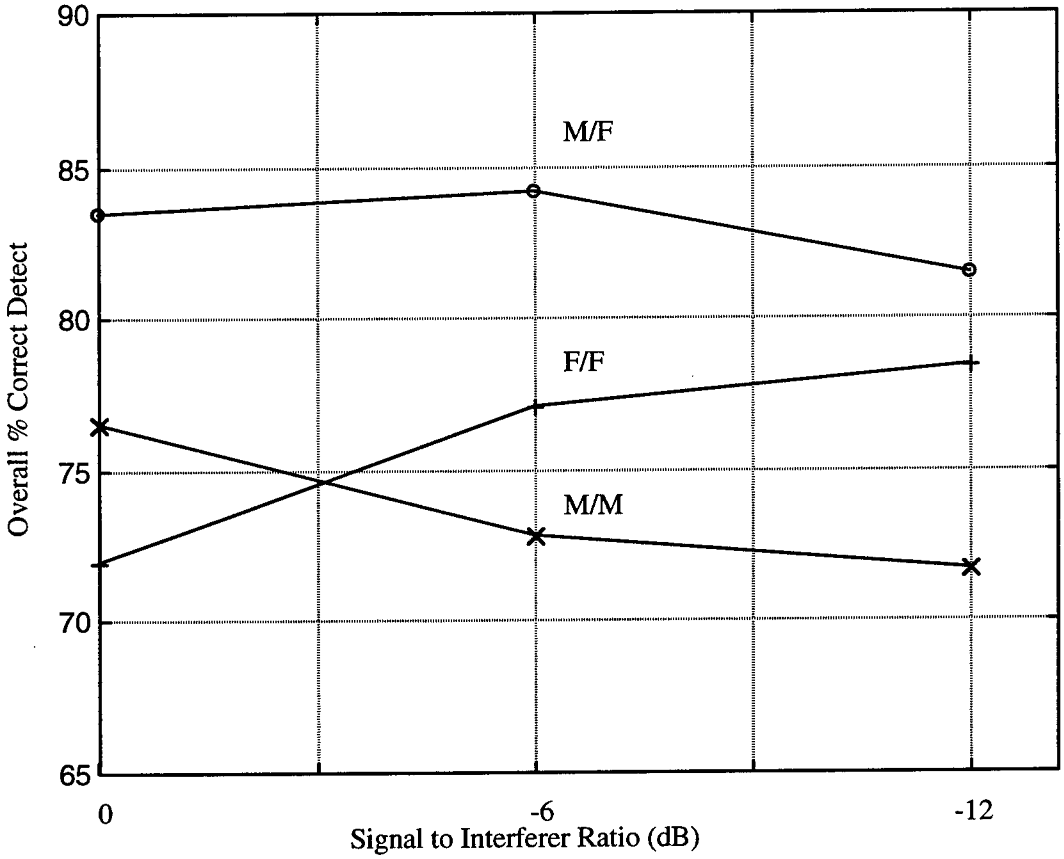


Figure 4.25: Plot of the overall percent correct detect of the VSDA on male/female (M/F), male/male (M/M) and female/female (F/F) speech mixtures at SIR = 0, -6, and -12 dB.

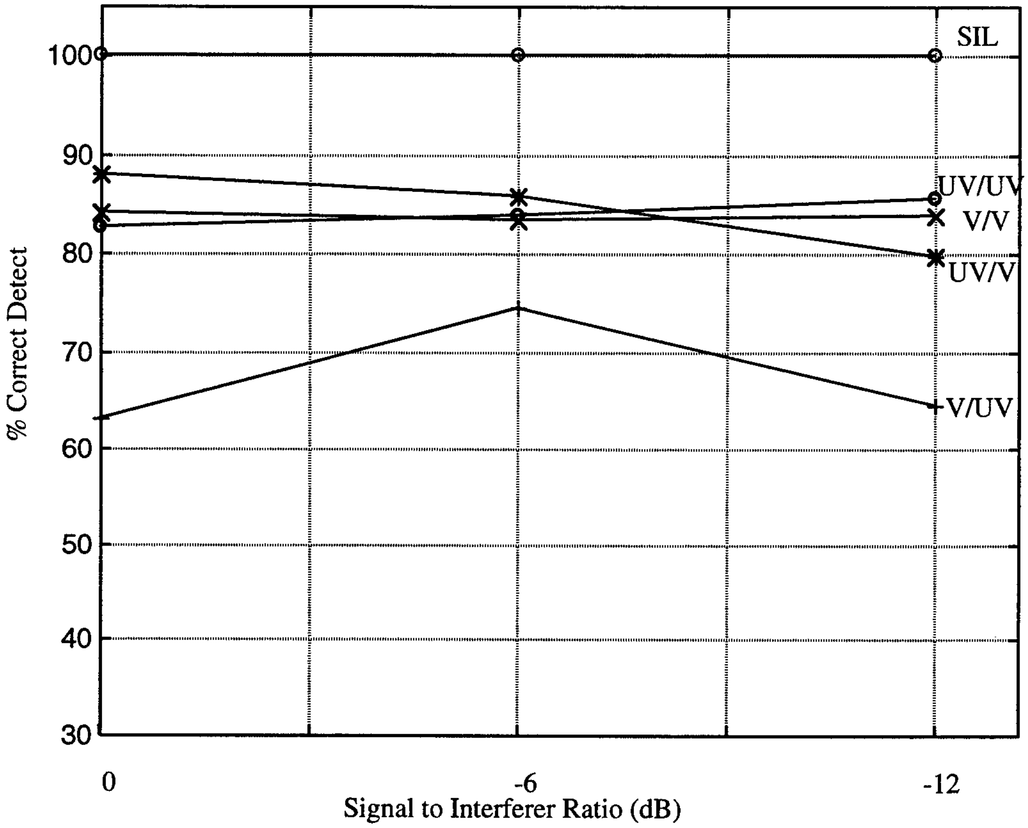


Figure 4.26: Plot of the cumulative percent correct detect of the five voicing states of co-channel speech in the male/female speech mixtures at SIR = 0, -6, and -12 dB.

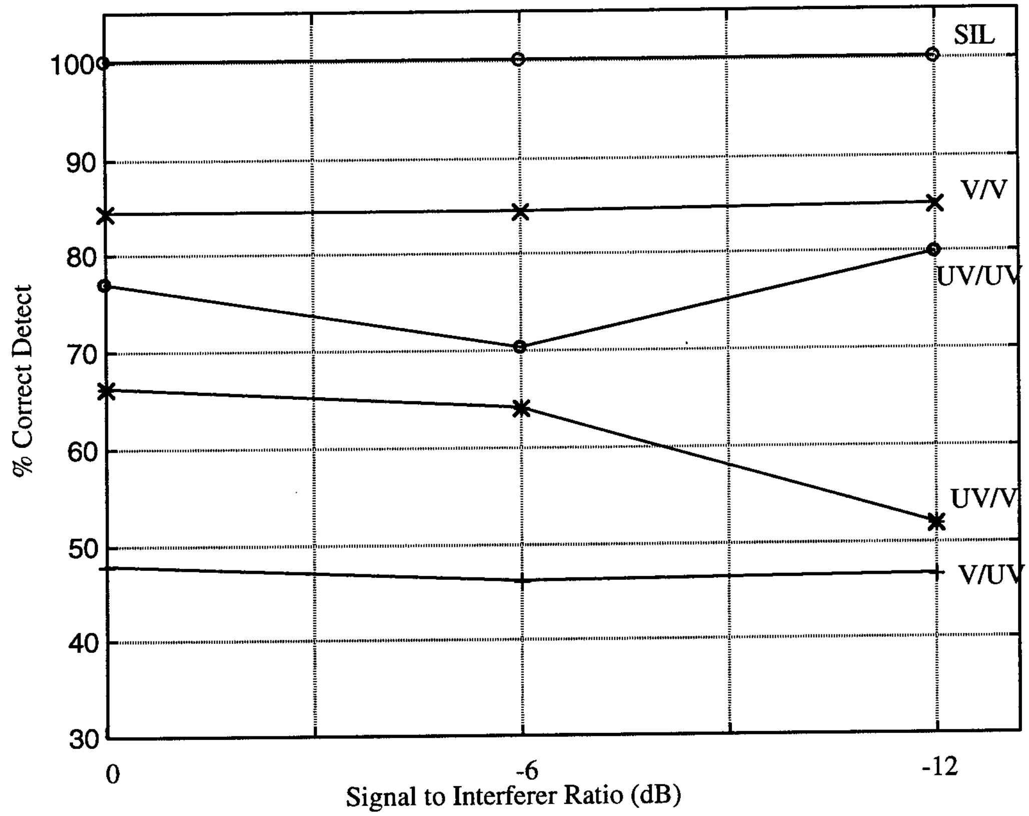


Figure 4.27: Plot of the cumulative percent correct detect of the five voicing states of co-channel speech in the male/male speech mixtures at SIR = 0, -6, and -12 dB.

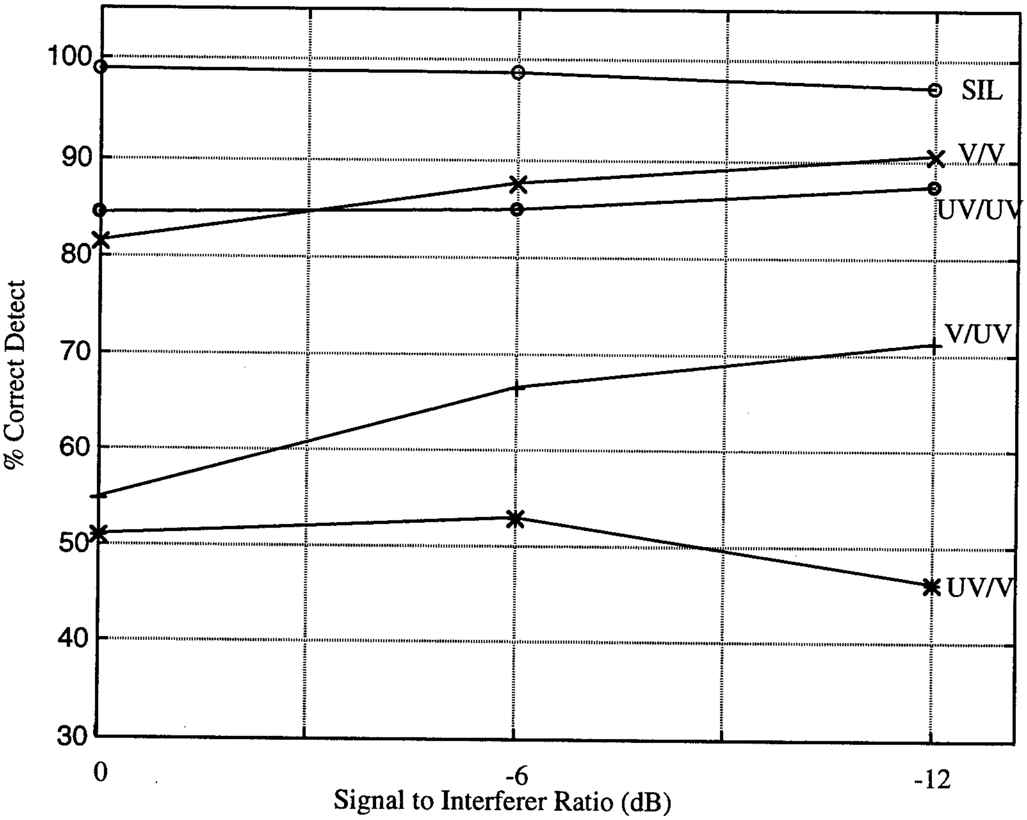


Figure 4.28: Plot of the cumulative percent correct detect of the five voicing states of co-channel speech in the female/female speech mixtures at SIR = 0, -6, and -12 dB.

Our last experiment was to test our co-channel speaker separation system on real speech signals. We used speech data taken from the RPI_COC database. The two speech signals came from a male and female speaker. The male talker is considered the desired speaker. Training for the voicing state determination algorithm was conducted using 60 seconds of speech for each speaker. This data was used to create 2000 segments of co-channel speech for each of the co-channel voicing states. This training is similar to the training using the TIMIT database speech. Testing was conducted on approximately six seconds worth of speech. The spoken sentence of the male talker was:

"A short segment of voiced speech can be modeled as a slowly-varying vocal tract filter."

The spoken sentence of the female talker was:

"Voicing state determination is a method of classifying the voicing state of a segment of speech."

The first test consisted of creating the co-channel speech signal by combining two independently recorded speech sentences. This provided an opportunity to test the VSDA (Bayes classifier) and the joint pitch estimation algorithm on the RPI_COC speech signals. The results of the VSDA are given in Table 4.30. We had an overall detection rate of 80.21%. We can see from this table that a significant percentage of the all-voiced speech segments (V/V) were misclassified as mixed voiced (V/UV, UV/V). However, there was only a small percentage of the V/UV and UV/V speech classified as V/V. The overall performance is consistent with the results obtained using the TIMIT_COC database, however for the individual co-channel voicing states, there is significant improvement in classifying the mixed voiced speech. This is in part due to the amount of training data used to train the voicing state classifier.

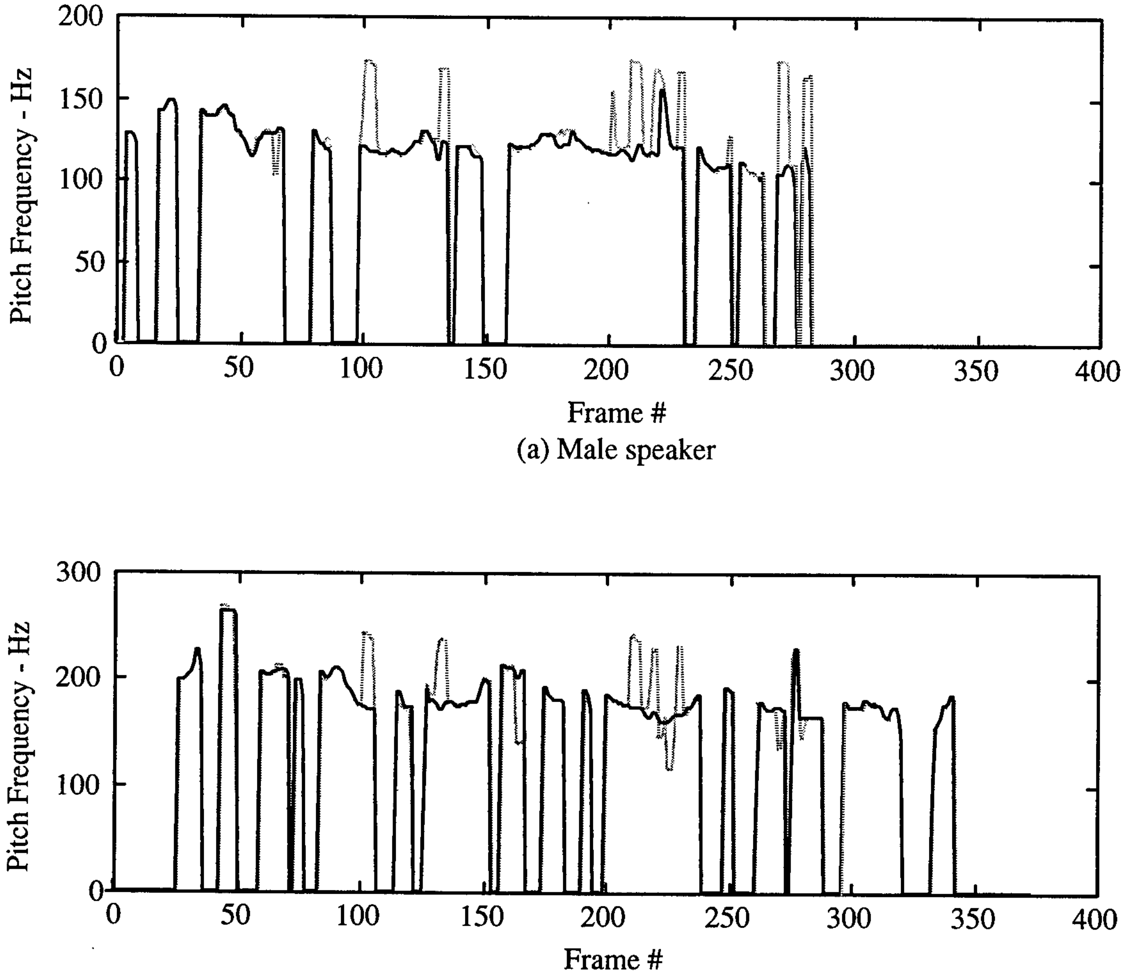
In Figure 4.29 we show a comparison between the pitch contours measured from the uncorrupted speech with the pitch contours measured from the co-channel speech. We obtained an accurate measurement of the pitch contour for the female speaker, but had several areas in which the pitch value of the male speaker measured from the co-channel speech signal was significantly higher than the referenced pitch value measured from the uncorrupted speech signal. This error was mainly due to poor harmonic suppression of the female talker.

In Figure 4.30 and Figure 4.31 a comparison between the original speech signal and the reconstructed speech signal obtained from our co-channel separation system is given. Here we can clearly identify the error which resulted from when the male speech signal was silent and the female talker was active. The relative error measurements were 0.4723, 0.4574, and 0.2202 for the male, female and co-channel signals respectively. These results are similar to the ones obtained with the TIMIT_COC database.

The second test consisted of creating a true co-channel speech signal by simultaneously recording the same two sentences given above, onto a single channel using a single microphone. The reconstructed speech signals from this test using our speaker separation system are provided in Figure 4.32. Listening tests comparing the reconstructed signals, obtained from the co-channel speech signal mixed on the computer, with the co-channel speech signal recorded from a single microphone demonstrate similar intelligibility performance.

Table 4.30: Confusion matrix, using the Bayes Classifier, of Male/Female co-channel speech mixture, from the RPI_COC database, SIR = 0 dB. Overall 80.21% of the speech segments were correctly classified. Values are in percent detection with raw scores in parentheses.

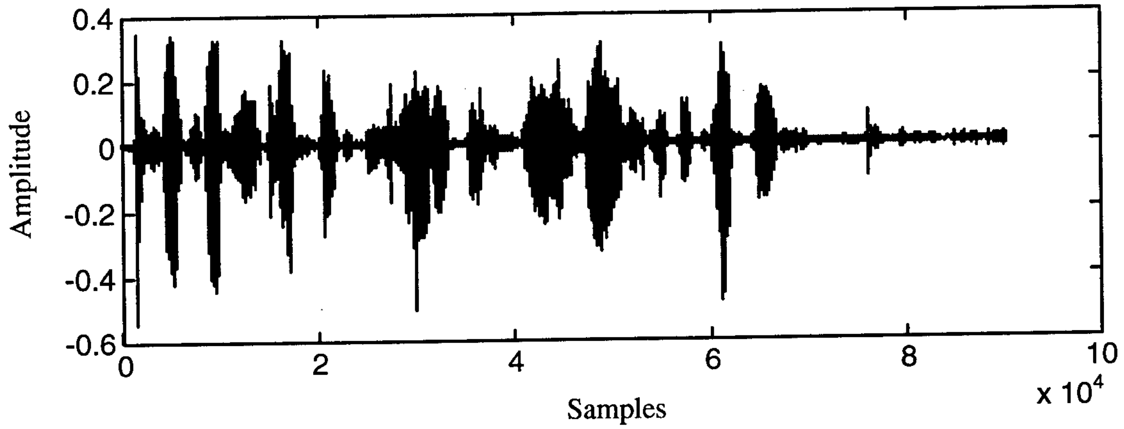
Voicing State	**SIL**	**V/V**	**V/UV**	**UV/V**	**UV/UV**
SIL	100 (36)	0	1.15 (1)	0	13.89 (5)
V/V	0	69.60 (87)	5.75 (5)	5.56 (5)	0
V/UV	0	12.00 (15)	78.16 (68)	1.11 (1)	2.78 (1)
UV/V	0	17.60 (22)	5.75 (5)	87.78 (79)	0
UV/UV	0	0.80 (1)	9.20 (8)	5.56 (5)	83.33 (30)



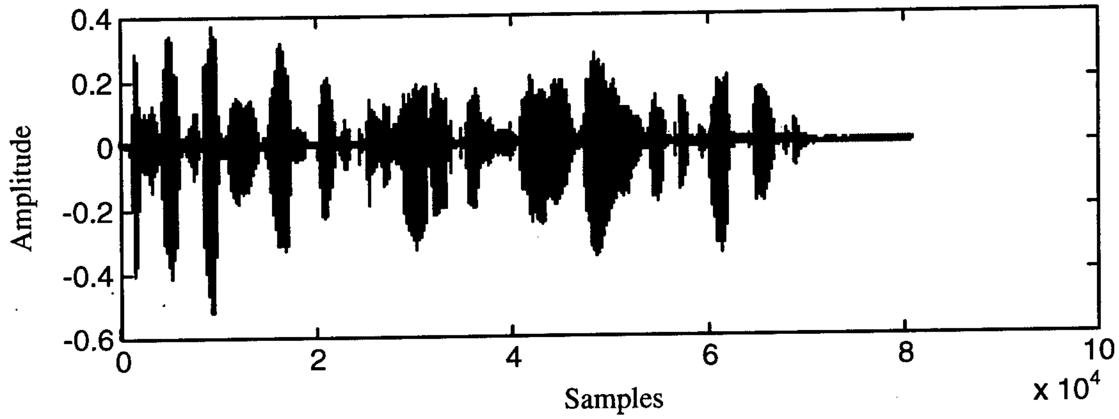
(a) Male speaker

(b) Female speaker

Figure 4.29: Comparison between pitch contour measured on uncorrupted speech (solid line) and from the co-channel speech (dotted line) for (a) male speaker, and (b) female speaker. Speech was taken from the RPI_COC database, SIR = 0 dB.

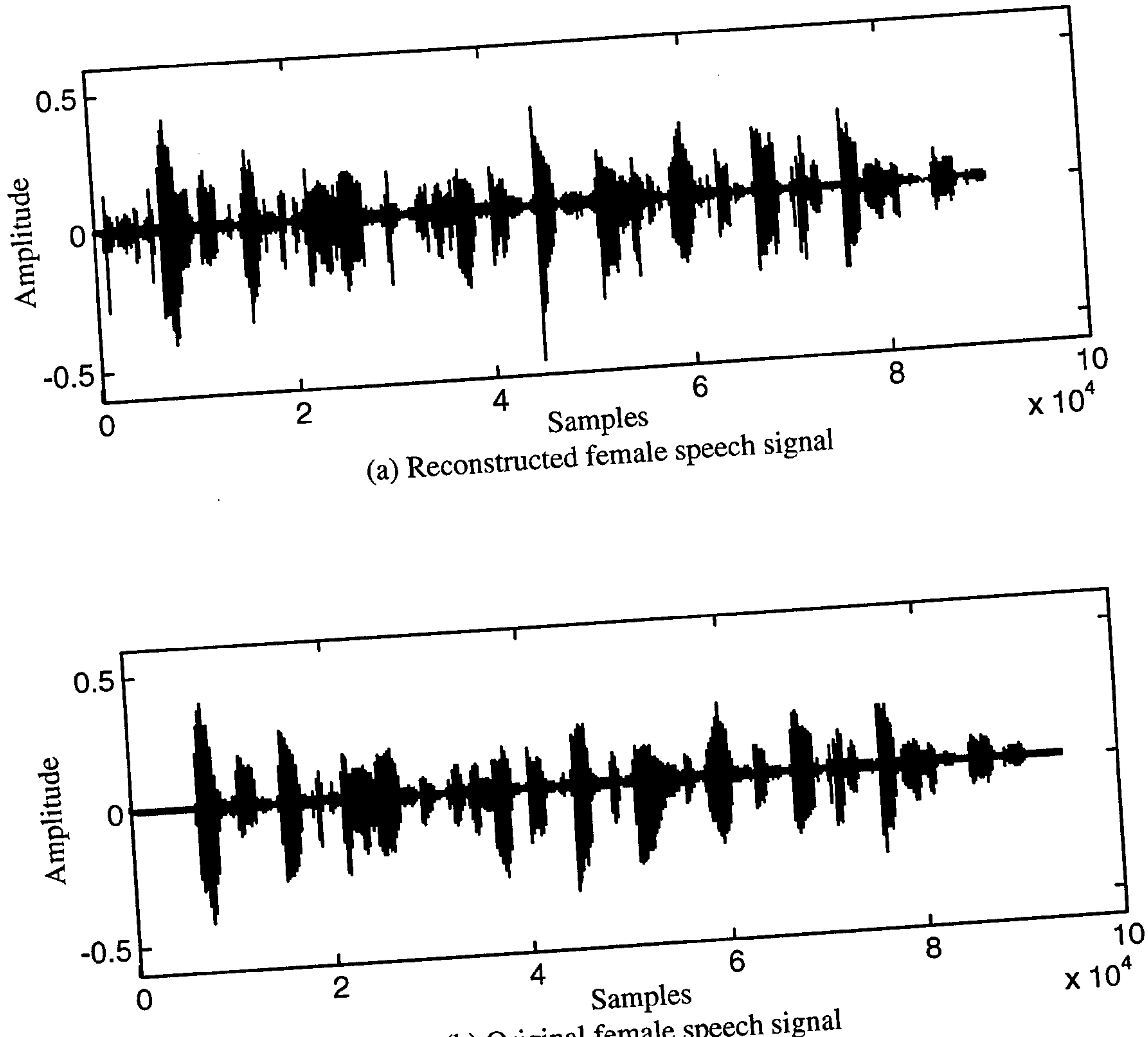


(a) Reconstructed male speech signal



(b) Original male speech signal

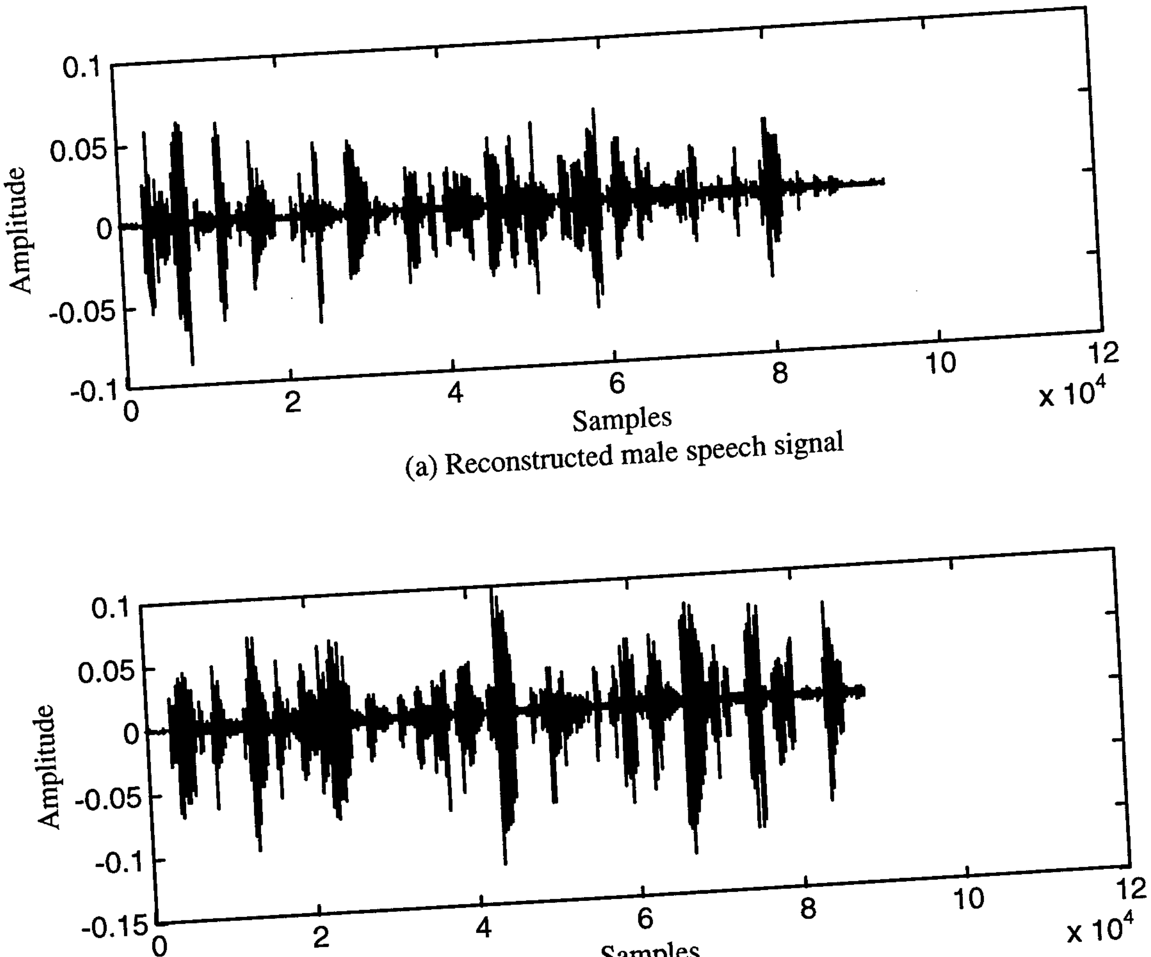
Figure 4.30: Male speech signal, from a male/female co-channel real speech mixture at SIR = 0 dB. (a) Reconstructed male speech signal and (b) original male speech signal.



(a) Reconstructed female speech signal

(b) Original female speech signal

Figure 4.31: Female speech signal, from a male/female co-channel speech mixture at SIR = 0 dB. (a) Reconstructed female speech signal and (b) original female speech signal.



(a) Reconstructed male speech signal

(b) Reconstructed female speech signal

Figure 4.32: Reconstructed speech signals from a true male/female co-channel speech signal with average SIR = 0 dB. (a) Reconstructed male speech signal and (b) reconstructed female speech signal.

5. DISCUSSION

5.1 Conclusions

In this research we have developed and presented a unique technique of separating overlapping speech signals recorded over a single channel. This system estimates the voicing state of each speaker, on a frame by frame basis. It then measures the pitch frequency of each segment of voiced speech and creates a pitch contour for each speaker. Using the pitch frequency, the system performs harmonic selection to assign initial estimates of the harmonics associated with each speaker. Our design uses a constrained nonlinear optimization algorithm to separate overlapping voiced speech signals and conventional filtering techniques to separate co-channel speech segments which are a combination of voiced and unvoiced speech.

We have made significant contributions to speaker separation. We developed a voicing state determination algorithm to classify the voicing state of co-channel speech. We have tested this algorithm using the Bayes, *k*-nearest neighbor, and Parzen window classifiers. We have tested and presented results for male/female, male/male, and female/female speech mixtures under varying SIRs. We have shown that we can accurately classify the voicing state of two overlapping speech signals. Performance on male/female speech mixtures is higher than same gender speech mixtures. We have developed a joint pitch estimation algorithm using the maximum likelihood pitch estimator and harmonic magnitude suppression. We have tested and presented results for varying speech mixtures under varying SIRs. This technique performs accurately when the pitch contours are well separated. We have developed a technique to perform harmonic selection to estimate the harmonic parameters for each speaker. We have developed and tested a constrained nonlinear least squares optimization algorithm to separate two overlapping voiced speech segments. We have shown that this technique can accurately estimate the amplitude, phase and frequency of multiple harmonics of overlapping vocalic speech segments. Finally, we have implemented and tested these algorithms into an end-to-end speaker separation system and have shown their performance in separating co-channel speech. Our system was tested and results were provided using real speech signals in simulated and real co-channel speaker environments.

Optimizing the amplitude and phase of each significant harmonic can be used to separate two overlapping voiced speech segments. Including the center frequency terms as optimization parameters significantly reduces the error between the measured co-channel signal and the sum of the estimated desired and interfering speech signals but significantly increases the processing time. However, this has a lesser effect on reducing the error between the true speech signals and the reconstructed speech signals. We attribute this to the error term used in the optimization routine. Obtaining a minimum error between the sum of the measured signals and the sum of the reconstructed signals does not guarantee a minimum error between the two measured signals and the two reconstructed signals.

We have demonstrated the capability to separate two speech signals recorded over a single channel. The results we obtained using speech extracted from the RPI_COC database provides a clear indication that our results using the TIMIT_COC database are an accurate representation of the overall performance we would expect in a co-channel speaker environment. Our voicing state determination algorithm, joint pitch estimation algorithm, and adaptive speaker separation system perform optimal with male/female speech mixtures at an average SIR = 0 dB.

5.2 Future Research

In our research we have presented a co-channel speaker separation system which does not use *a priori* information to process and separate co-channel speech. Our system represents a significant contribution to the area of co-channel speaker separation. However, this technique brings to the forefront several areas which still need further investigation.

5.2.1 Error Measurement

Separating the speech signals based on finding the minimum error between the sum of the measured speech signals and the sum of the estimated speech signals will not guarantee a minimum error between the true speech signals and the estimated speech signals. Other error terms need to be investigated.

5.2.2 Voicing State Determination

A technique needs to be developed which will perform unsupervised learning of the VSDA using training data extracted from the co-channel speech. This would then require a technique which can identify the intervals of time when one speaker is active and the other speaker is silent. In a realistic co-channel speaker environment, one speaker may be silent for extended intervals of time, thereby providing the necessary training data and limiting separation to shorter segments of speech. This will improve the robustness of our co-channel speaker system to operate in more realistic co-channel speaker environments.

5.2.3 Joint Pitch Estimation

Accurate pitch contours of both speakers are crucial to most speaker separation systems. Improved joint pitch estimation and pitch tracking of co-channel speech is required to advance the current performance of most co-channel speaker separation systems.

BIBLIOGRAPHY

[1] O.M.M. Mitchell, C.A. Ross, and G.H. Yates, "Signal Processing for a Cocktail Party Effect," J. Acoust. Soc. Am., 50(2):656-560, October 1971.

[2] T.W. Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection," *J. Acoust. Soc. Am.*, 60(4):911-918, October 1976.

[3] T.W. Parsons and M.R. Weiss, "Enhancing intelligibility of speech in Noisy or Multi-Talker Environments," Rome Air Development Center, Griffiss Air Force Base, NY, Technical Report RADC-TR-75-155, June 1975.

[4] T.W. Parsons, "Study and Development of Speech Separation Techniques," Rome Air Development Center, Griffiss Air Force Base, NY, Technical Report RADC-TR-78-105, May 1978.

[5] T.W. Parsons, "Multi-Talker Separation," Rome Air Development Center, Griffiss Air Force Base, NY, Technical Report RADC-TR-79-242, October 1979.

[6] V.C. Shields, Jr., "Separation of Added Speech Signals by Digital Comb Filtering," Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA, September 1970.

[7] R.H. Frazier, "An Adaptive Filtering Approach toward Speech Enhancement," Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA, June 1975.

[8] J.K. Everton, Sr., "The Separation of the Voice Signals of Simultaneous Speakers," Ph.D. Thesis, University of Utah, Salt Lake City, Utah, June 1975.

[9] R.J. Dick, "Cochannel Interference Separation," Rome Air Development Center, Griffiss Air Force Base, NY, Technical Report RADC-TR-80-365, December 1980.

[10] B.A. Hanson and D.Y. Wong, "Processing Techniques for Intelligibility Improvement to Speech with Cochannel Interference," Rome Air Development Center, Griffiss Air Force Base, NY, Technical Report RADC-TR-83-225, September 1983.

[11] B.A. Hanson, D.Y. Wong, and B.H. Juang, "Speech Enhancement with Harmonic Synthesis," Intl. Conf. on Acoust. Speech and Signal Process., Boston, MA, pp. 1122-1125, April 1983.

[12] B.A. Hanson and D.Y. Wong, "The Harmonic Magnitude Suppression (HMS) Technique for Intelligibility Enhancement in the Presence of Interfering Speech," Intl. Conf. on Acoust. Speech and Signal Process., San Diego, CA, pp. 18A.5.1-18A.5.4, March 1984.

[13] C.K. Lee and D.G. Childers, "Cochannel Speech Separation:, *J. Acoust. Soc. Am.*, 83(1):274-280, January 1988.

[14] M. Weintraub, "A Computational Model for Separating Two Simultaneous Talkers," Intl. Conf. on Acoust. Speech and Signal Process., pp. 80-84, April 1986.

[15] S.T. Alexander, "Adaptive Reduction of Interfering Speaker Noise using the Least Mean Squares Algorithm," Intl. Conf. on Acoust. Speech and Signal Process., pp. 728-731, Tampa FL, March 1985.

[16] J.A. Naylor and S.F. Boll, "Techniques for Suppression of an Interfering Talker in Cochannel Speech," Intl. Conf. on Acoust. Speech and Signal Process., pp. 205-208, Dallas TX, April 1987.

* [17] J.A. Naylor, "Interference Reduction Model," Rome Air Development Center, Griffiss Air Force Base, NY, Technical Report RADC-TR-87-175, October 1987.

[18] G.E. Kopec and M.A. Bush, "An LPC-based Spectral Similarity Measure for Speech Recognition in the Presence of Co-channel Speech Interference," Intl. Conf. on Acoust. Speech and Signal Process., pp. 270-273, Glasgow, Scotland, May 1989.

[19] C. Rogers, D. Chien, M. Featherston, and K. Min, "Neural Network Enhancement for a Two Speaker Separation System," Intl. Conf. on Acoust. Speech and Signal Process., pp. 357-360, Glasgow, Scotland, May 1989.

[20] A.P. Varga and R.K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," Intl. Conf. on Acoust. Speech and Signal Process., pp. 845-848, Albuquerque, NM, April 1990.

[21] H. Gish, M. Siu, and R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification," Intl. Conf. on Acoust. Speech and Signal Process., pp. 873-876, Toronto, Canada, 1991.

[22] Y.H. Gu and W.M.G. van Bokhoven, "Co-channel Speech Separation using Frequency Bin Nonlinear Adaptive Filtering," Intl. Conf. on Acoust. Speech and Signal Process., pp. 949-952, Toronto, Canada, 1991.

* Distr authorized to US gov't agencies and their contractors, Oct 87.

[23] T.F. Quatieri and R.G. Danisewicz, “An Approach to Co-channel talker Interference Suppression Using a Sinusoidal Model for Speech,” *IEEE Trans. on Acoust. Speech and Signal Processing,* 38(1):56-69, January 1990.

* [24] J. Naylor and J. Porter,” Cochannel EDM,” Rome Laboratory, Griffiss Air Force Base, NY, Technical Report RL-TR-93-72, May 1993.

[25] J. Naylor and J. Porter, “An Effective Speech Separation System Which Requires No *A Priori* Information,” Intl. Conf. on Acoust. Speech and Signal Process., Toronto, Canada, 1991.

[26] M.A. Zissman, “Cochannel Talker Interference Suppression,” Lincoln Laboratory, Lexington, MA, Technical Report-895, July 1991.

[27] M.A. Zissman, C.J. Weinstein, L.D. Braida, R.M. Uchanski and W.M. Rabinowitz, “Speech-State-Adaptive Simulation of Cochannel Talker Interference Suppression,” Intl. Conf. on Acoust. Speech and Signal Process., pp. 361-364, Glasgow, Scotland, May 1989.

[28] M. Savic, H. Gao and J.S. Sorensen, “Co-channel Speaker Separation Based on Maximum-Likelihood Deconvolution,” Speech Research Symposium XI, 1993.

[29] D. P. Morgan, E.B. George, L.T. Lee and S.M. Kay, “Co-channel Speaker Separation,” Intl. Conf. on Acoust. Speech and Signal Process., 1995.

[30] J.D. Wise, J.R. Caprio and T.W. Parks, “Maximum Likelihood Pitch Estimation,” *IEEE Trans. on Acoust., Speech and Signal Processing*,” vol. 24(5), pp. 418-423, October 1976.

*Distr authorized to US gov't agencies and their contractors; May 93.

[31] D. P. Morgan, E.B. George, L.T. Lee and S.M. Kay, "Co-channel Speaker Separation," *IEEE Trans. on Acoust., Speech and Signal Processing.*, 1997.

[32] P. Margos, J.F. Kaiser and T. Quatieri, "Energy Separation in Signal Modulations with Applications to Speech Analysis," *IEEE Trans. on Signal Processing,"* vol. 41(10), pp. 3024-3051, October 1993.

[33] T. Parsons, *Voice and Speech Processing*, McGraw-Hill Book Company, New York, 1987.

[34] B.S. Atal and L.R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. ASSP-24(3), June 1976.

[35] L.R. Rabiner and M.R. Sambur, "Application of an LPC Distance Measure to the Voiced-Unvoiced-Silence Detection Problem," *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. ASSP-25(4), August 1977.

[36] L.J. Siegal and A.C. Bessey, "Voiced/Unvoiced/Mixed Excitation Classification of Speech," *IEEE Trans. on Acoust., Speech and Signal Processing,* vol. ASSP-30(3), June 1982.

[37] H.C. Woodsum, M.J. Pitaro and S.M. Kay, "An Iterative Algorithm for the Simultaneous Estimation of Pitch from Two Interfering Speech Signals," Proc. of the Chatham Digital Signal Processing Workshop, pp. 5.6.1-5.6.2, October 1986.

[38] M.R. Schroeder, "Models of Hearing," *Proceedings of the IEEE,"* vol. 63(9), pp. 1332-1350, September 1975.

[39] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[40] D. O'Shaughnessy, *Speech Communication*, Addison-Wesley Publishing Company, Reading MA, 1987.

[41] S.L. Marple, *Digital Spectral Analysis with Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[42] R.J. McAulay and T.F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. ASSP-34(4), August 1986.

[43] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer-Verlag, Berlin, 1983.

[44] Y.M. Perlmutter, L.D. Braida, R.H. Frazier and A.V. Oppenheim, "Evaluation of a Speech Enhancement System," Intl. Conf. on Acoust. Speech and Signal Process., pp. 212-215, May 1977.

[45] E.C. Cherry and R. Wiley, "Speech Communications in Very Noisy Environments," *Nature*, vol. 214, pp. 1164, 1967.

[46] S. Furui, *Digital Speech Processing, Synthesis and Recognition*," Marcel Dekker, Inc., New York, 1989.

[47] E.M.L. Beale, *Introduction to Optimization*," John Wiley & Sons, New York, 1988.

[48] T.R. Cuthbert, Jr., *Optimization using Personal Computers: With Applications to Electrical Networks*," John Wiley & Sons, New York, 1987.

[49] D. Jacobs, *The State of the Art in Numerical Analysis*," Academic Press, New York, 1977.

[50] R.V. Churchill and James W. Brown, *Complex Variables and Applications 5th ed.,"* McGraw-Hill, Inc., New York, NY, 1990.

[51] H. L. Van Trees, *Detection, Estimation, and Modulation Theory Part I,* John Wiley & Sons, New York, NY, 1968.

[52] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, NY, 1973.

[53] K. Fukunaga and P. M. Narendra, "A Branch and Bound Algorithm for Computing the k-Nearest Neighbors," *IEEE Transactions on Computers*, pp. 750-753, July 1975.

[54] L. Rabiner and B-H Juang, *Fundamentals of Speech Recognition,* Prentice Hall, Englewood Cliffs, NJ, 1993.

☆U.S. GOVERNMENT PRINTING OFFICE: 1999-510-079-81216